IFIP AICT 443

Christian Pötzsche
Clemens Heuberger
Barbara Kaltenbacher
Franz Rendl
(Eds.)

System Modeling and Optimization

26th IFIP TC 7 Conference, CSMO 2013
Klagenfurt, Austria, September 9–13, 2013
Revised Selected Papers

Springer

IFIP Advances in Information and Communication Technology 443

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 IFIPWorld 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.

More information about this series at http://www.springer.com/series/6102

Christian Pötzsche · Clemens Heuberger
Barbara Kaltenbacher · Franz Rendl (Eds.)

System Modeling and Optimization

26th IFIP TC 7 Conference, CSMO 2013 Klagenfurt, Austria, September 9–13, 2013 Revised Selected Papers

Springer

Editors
Christian Pötzsche
Institut für Mathematik
Alpen-Adria-Universität Klagenfurt
Klagenfurt
Austria

Clemens Heuberger
Institut für Mathematik
Alpen-Adria-Universität Klagenfurt
Klagenfurt
Austria

Barbara Kaltenbacher
Institut für Mathematik
Alpen-Adria-Universität Klagenfurt
Klagenfurt
Austria

Franz Rendl
Institut für Mathematik
Alpen-Adria-Universität Klagenfurt
Klagenfurt
Austria

ISSN 1868-4238 ISSN 1868-422X (electronic)
IFIP Advances in Information and Communication Technology
ISBN 978-3-662-45503-6 ISBN 978-3-662-45504-3 (eBook)
DOI 10.1007/978-3-662-45504-3

Library of Congress Control Number: 2014956237

Springer Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer-Verlag GmbH Berlin Heidelberg is part of Springer Science+Business Media (www.springer.com)

Preface

Every 2 years, the International Federation for Information Processing Technical Committee 7 (IFIP TC 7)—System Modeling and Optimization—arranges highly regarded conferences on several topics of Applied Optimization, such as Optimal Control of Ordinary and Partial Differential Equations, Modeling and Simulation, Inverse Problems, Nonlinear, Discrete, and Stochastic Optimization, as well as Industrial Applications.

The collection in your hands contains selected papers presented at the 26th IFIP TC 7 Conference held at the Alpen-Adria-Universität Klagenfurt, Austria during September 8–13, 2013,

http://ifip2013.uni-klu.ac.at

The conference was organized by the local Institute of Mathematics. Preceding conferences in this series were held in Berlin, Germany (2011), Buenos Aires, Argentina (2009), Cracow, Poland (2007), and Turin, Italy (2005). The scientific program of the 2013 event consisted of 10 plenary talks, 33 minisymposia (from 4–12 invited talks each), and 5 contributed sessions. In total, this resulted in 235 talks. Altogether 268 participants from 32 countries came to the conference, the largest groups with respect to country of origin were Germany (80), Austria (66), Poland (21), France (15), UK (11), USA (11), Romania (10), Russia (10), Italy (8), Belgium (5), and the Czech Republic (5).

The 34 refereed contributions to these proceedings cover the latest progress in a wide range of topics discussed at the meeting.

Acknowledgments: We are grateful to the sponsors, namely the European Science Foundation (ESF), Europäisches Patentamt, Die Kärntner Sparkasse, the Land Kärnten, and the city of Klagenfurt.

Finally, we would like to thank our referees.

July 2014

Christian Pötzsche
Clemens Heuberger
Barbara Kaltenbacher
Franz Rendl

Organization

Committees

International Scientific Committee

Jacques Henry (Chair)	Inria/Université Bordeaux 1, France
Arun Bagchi	University of Twente, The Netherlands
A.V. Balakrishnan	University of California, Los Angeles, USA
Héctor Cancela	Universidad de la República, Uruguay
J. Valério Carvalho	Universidade do Minho, Portugal
Gianni Di Pillo	Università di Roma "La Sapienza", Italy
Yu.G. Evtushenko	Russian Academy of Sciences, Moscow, Russia
Janusz Granat	Warsaw Institute of Technology, Poland
Alfred Kalliauer	VERBUND – Austrian Power Trading, Austria
Peter Kall	University of Zurich, Switzerland
Hisao Kameda	University of Tsukuba, Japan
Irena Lasiecka	University of Virginia, USA
István Maros	University of Pannonia, Hungary
Kurt Marti	Federal Armed Forces University, Germany
Lukasz Stettner	Polish Academy of Sciences, Poland
Fredi Tröltzsch	TU Berlin, Germany
Lidija Zadnik-Stirn	Univerza v Ljubljani, Slovenia
Jean-Paul Zolesio	CNRS and Inria, Sophia Antipolis, France

Local Scientific Committee

Barbara Kaltenbacher	Alpen-Adria-Universität Klagenfurt, Austria
Clemens Heuberger	Alpen-Adria-Universität Klagenfurt, Austria
Christian Pötzsche	Alpen-Adria-Universität Klagenfurt, Austria
Franz Rendl	Alpen-Adria-Universität Klagenfurt, Austria

Participants

Contents

Stochastic Maximum Principle for Hilbert Space Valued Forward-Backward Doubly SDEs with Poisson Jumps

AbdulRahman Al-Hussein[1(✉)] and Boulakhras Gherbal[2]

[1] Department of Mathematics, College of Science, Qassim University, P.O. Box 6644, Buraydah 51452, Saudi Arabia
alhusseinqu@hotmail.com, hsien@qu.edu.sa
[2] Laboratory of Applied Mathematics, University of Mohamed Khider, P.O. Box 145, 07000 Biskra, Algeria

Abstract. In this paper we study the stochastic maximum principle for a control problem in infinite dimensions. This problem is governed by a fully coupled forward-backward doubly stochastic differential equation (FBDSDE) driven by two cylindrical Wiener processes on separable Hilbert spaces and a Poisson random measure. We allow the control variable to enter in all coefficients appearing in this system.

Existence and uniqueness of the solutions of FBDSDEs and an extended martingale representation theorem are provided as well.

Keywords: Wiener process · Poisson process · Forward-backward doubly stochastic differential equation · Maximum principle

1 Introduction

Backward stochastic differential equations in infinite dimensions (BSDEs) were studied by Hu and Peng in [6], Tessitore in [17] and Al-Hussein in [3]. Al-Hussein proved in [3] the existence and uniqueness of the solutions to BSDEs in infinite dimensions driven by genuine Q-Wiener processes (and also cylindrical Wiener processes) on separable Hilbert spaces. He gave also a representation of the solution of a system of semi-linear parabolic PDEs and found viscosity solutions to such PDEs. In [4] sufficient conditions of optimality for backward stochastic evolution equations on Hilbert spaces are derived. Several references in these directions are recorded in [4]. These works give a motivational base to study the maximum principle for optimality of forward-backward stochastic differential equations (FBSDEs) in infinite dimensions. In fact, Yin and Wang [19], proved the existence and uniqueness of the solutions of FBSDEs with Poisson jumps in Hilbert space and with bounded random terminal times. Their work relies on

A. Al-Hussein—This work is supported by the Science College Research Center at Qassim University, project no. SR-D-012-1958.

B. Gherbal—It is also supported by the Algerian PNR project no. 8/u 07/857.

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 1–10, 2014.
DOI: 10.1007/978-3-662-45504-3_1

those in [16] and the method of continuation given in [7]. Developing applications to such FBSDEs as for example in [3] are not yet well studied.

Let us now talk about more general equations. In finite dimensions, a fully coupled forward-backward doubly stochastic differential equation (FBDSDE) was introduced by Peng and Shi in [12]. Such equations are generalizations of stochastic Hamilton systems. Al-Hussein and Gherbal in [5] studied a stochastic control problem governed by a fully coupled multi-dimensional FBDSDE with Poisson jumps.

In the present work, we shall work in infinite dimensions and try to derive the stochastic maximum principle for optimal control of fully coupled FBDSDEs with jumps; see (1) below. Moreover, existence and uniqueness of the solutions to infinite dimensional FBDSDEs along with an extended martingale representation theorem will be provided as well.

Applications of such equations can be gleaned from [5]. Our formulation of these equations as well as cost functionals are given in abstract forms to allow the possibility to work directly in the case of partial information on one hand and on the other hand to cover most of the applications available in the literature. For instance, a linear quadratic case can be given as a concrete and useful example. For more details of this example, we refer the reader to [15] or [18]. In fact, many applications of FBDSDE either in finance or to stochastic PDEs can be developed in parallel to those provided in the literature.

Our results here can be generalized easily to the case of a stochastic relaxed control problem governed by a nonlinear fully coupled FBDSDE with Poisson jumps, which involves relaxed controls. We refer the reader to Ahmed et al. [1], in this respect.

The paper is organized as follows. Notation and an extended martingale representation theorem are recorded in Sect. 2. Section 3 is devoted to stating the stochastic optimal control problem, which is governed by FBDSDE (1). Existence and uniqueness of the solutions of FBDSDEs are included in Sect. 4. Finally, in Sect. 5 we establish the stochastic maximum principle of our control problem.

2 Notation and an Extended Martingale Representation Theorem

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space. Let H_1 and H_2 be two separable Hilbert spaces. Assume that $(W_t)_{t\in[0,T]}$ and $(B_t)_{t\in[0,T]}$ are two cylindrical Wiener processes on H_1 and H_2 respectively, where T is a fixed positive number. Let η be a Poisson point process with values in a measurable space $(\Theta, \mathcal{B}(\Theta))$. We denote by $\nu(d\theta)$ to the characteristic measure of η, which is assumed to be a σ-finite measure on $(\Theta, \mathcal{B}(\Theta))$, by $N(d\theta, dt)$ to the Poisson counting measure induced by η with compensator $\nu(d\theta)dt$, and by $\tilde{N}(d\theta, dt) = N(d\theta, dt) - \nu(d\theta)dt$ to the compensation of the jump measure $N(\cdot, \cdot)$ of η. We assume that the three processes B, W and η are mutually independent.

For each $t \in [0, T]$, define

$$\mathcal{F}_t := \mathcal{F}_t^W \vee \mathcal{F}_{t,T}^B \vee \mathcal{F}_t^\eta,$$

where

$$\mathcal{F}_t^W := \sigma\{l(W_s) :\ 0 \le s \le t, l \in H_1^*\} \vee \mathcal{N},$$
$$\mathcal{F}_{t,T}^B := \sigma\{l(B_r) - l(B_t) :\ t \le r \le T, l \in H_2^*\},$$
$$\mathcal{F}_t^\eta := \sigma\{\eta_s :\ 0 \le s \le t\} \vee \mathcal{N},$$

and $\mathcal{N}$ is the collection of all $\mathbb{P}$-null sets of $\mathcal{F}$.

Note that $\{\mathcal{F}_t\}_{t\in[0,T]}$ does not constitute a filtration because it is not increasing nor decreasing.

Let us set the following spaces of solutions.

For a separable Hilbert space E, let $\mathcal{M}^2(0,T;E)$ denote the set of jointly measurable processes $\{\mathcal{Y}_t, t \in [0,T]\}$ taking values in E, and satisfy: $\mathcal{Y}_t$ is $\mathcal{F}_t$-measurable for a.e. $t \in [0,T]$, and

$$\mathbb{E}\big[\int_0^T |\mathcal{Y}_t|_E^2\, dt\big] < \infty.$$

Let $L_\nu^2(E)$ be the set of $\mathcal{B}(\Theta)$-measurable mapping k with values in K such that

$$|||k||| := \big[\int_\Theta |k(\theta)|_E^2\, \nu(d\theta)\big]^{\frac{1}{2}} < \infty.$$

Denote by $\mathcal{V}_\eta^2(0,T;E)$ to the set of processes $\{\mathfrak{K}_t,\ t \in [0,T]\}$ that take their values in $L_\nu^2(K)$ and satisfy: $\mathfrak{K}_t$ is $\mathcal{F}_t$-measurable for a.e. $t \in [0,T]$, and

$$\mathbb{E}\big[\int_0^T \int_\Theta |\mathfrak{K}_t(\theta)|_E^2\, \nu(d\theta)dt\big] < \infty.$$

Finally, fixing a fixed separable Hilbert space K, we set

$$\mathbb{M}^2 := \mathcal{M}^2(0,T;K) \times \mathcal{M}^2(0,T;K) \ \times \mathcal{M}^2(0,T;L_2(H_2,K)) \\ \times \mathcal{M}^2(0,T;L_2(H_1,K)) \times \mathcal{V}_\eta^2(0,T;K).$$

Here $L_2(E,K)$ denotes the space of all Hilbert-Schmidt operators from E into K, for $E = H_1, H_2$, with inner product dented by $||\cdot||$. Then $\mathbb{M}^2$ is a Hilbert space with respect to the norm $\|\cdot\|_{\mathbb{M}^2}$ given, for $\Lambda_\cdot = (x_\cdot, Y_\cdot, z_\cdot, Z_\cdot, \xi_\cdot)$, by

$$\|\Lambda_\cdot\|_{\mathbb{M}^2}^2 \\ := \mathbb{E}\big[\int_0^T |x_t|^2\, dt + \int_0^T |Y_t|^2\, dt + \int_0^T \|z_t\|^2\, dt + \int_0^T \|Z_t\|^2\, dt + \int_0^T |||\xi_t|||^2\, dt\big].$$

We close this section by providing an extended martingale representation theorem.

Theorem 1. *Let ρ and g be elements of $L^2(\Omega, \mathcal{F}_T, \mathbb{P}; K)$ and $\mathcal{M}^2(0,T;L_2(H_1,K))$, respectively. If M is the martingale*

$$M(t) = \mathbb{E}\,[\,\rho + \int_0^T g(t)\, \overleftarrow{dB}_t \,|\, \mathcal{E}_t\,],\ 0 \le t \le T,$$

where $\mathcal{E}_t := \mathcal{F}_t^W \vee \mathcal{F}_T^B \vee \mathcal{F}_t^\eta$, *then there exist unique elements* (ϕ, κ) *of* $\mathcal{M}^2(0,T; L_2(H_2,K)) \times \mathcal{V}_\eta^2(0,T;K)$ *such that*

$$M(t) = M(0) + \int_0^t \phi_s \, dW_s + \int_0^t \int_\Theta \kappa_s(\theta) \tilde{N}(d\theta, ds).$$

Here the integral with respect to $\overleftarrow{dB}$ is a backward Itô integral, while the integral with respect to dW is a standard forward Itô integral.

This result is known in finite dimensions (i.e. when all Hilbert spaces are Euclidean spaces), as it can be seen easily by combining the well known martingale representation theorem for Brownian motions and Poisson random measure (e.g. see [10, Lemma 4.2] or [8]) and [11, Proposition 1.2]. The proof of this infinite dimensional version can be gleaned by mimicking the ideas of proofs in [11, Proposition 1.2], [2, Theorem 3.1] and [10, Lemma 4.2].

Such a theorem is in fact essential for finding solutions to BDSDEs and decoupled (or coupled) FBDSDEs; see e.g. [9] for using martingale representation theorem to show the existence of solutions to FBSDEs in continuous situations.

3 Statement of the Control Problem

Let $\mathcal{O}$ be a separable Hilbert space and U be a nonempty convex of $\mathcal{O}$. We say that $v. : [0,T] \times \Omega \to \mathcal{O}$ is *admissible* if $v. \in \mathcal{M}^2(0,T;\mathcal{O})$ and $v_t \in U$ *a.e.* t, *a.s.* The set of admissible controls will be denoted by $\mathcal{U}_{ad}$. Consider the following controlled K-valued fully coupled FBDSDE with jumps:

$$\begin{cases} dx_t = B(t, x_t, Y_t, z_t, Z_t, \xi_t, v_t)dt + \Sigma(t, x_t, Y_t, z_t, Z_t, \xi_t, v_t)dW_t \\ \qquad\qquad + \int_\Theta \Phi(t, x_t, Y_t, z_t, Z_t, \xi_t, v_t, \theta)\tilde{N}(d\theta, dt) - z_t \overleftarrow{dB}_t, \\ dY_t = -F(t, x_t, Y_t, z_t, Z_t, \xi_t, v_t)dt - G(t, x_t, Y_t, z_t, Z_t, \xi_t, v_t)\overleftarrow{dB}_t \\ \qquad\qquad + Z_t dW_t + \int_\Theta \xi_t(\theta)\tilde{N}(d\theta, dt), \\ x_0 = \pi \in K, Y_T = h(x_T), \ t \in (0,T), \end{cases} \tag{1}$$

where the coefficients

$$\begin{aligned} &B, F : \Omega \times [0,T] \times K \times K \times L_2(H_2;K) \times L_2(H_1;K) \times L_\nu^2(K) \times \mathcal{O} \to K, \\ &\Sigma, G : \Omega \times [0,T] \times K \times K \times L_2(H_2;K) \times L_2(H_1;K) \times L_\nu^2(K) \times \mathcal{O} \to L_2(H_1;K), \\ &\Phi : \Omega \times [0,T] \times K \times K \times L_2(H_2;K) \times L_2(H_1;K) \times L_\nu^2(K) \times \mathcal{O} \times \Theta \to K, \\ &h : \Omega \times K \to K, \end{aligned}$$

are measurable and $v. \in \mathcal{U}_{ad}$. More conditions will be assumed in Sect. 3. The mapping h is defined, for $(\omega, x) \in \Omega \times K$, by $h(\omega, x) := c\,x + \zeta(\omega)$, where $c \neq 0$ is a constant and ζ is a fixed arbitrary element of $L^2(\Omega, \mathcal{F}_T, \mathbb{P}; K)$.

Definition 1. *A solution of (1) is a quintuple* (x, Y, z, Z, ξ) *of stochastic processes such that* (x, Y, z, Z, ξ) *belongs to* $\mathbb{M}^2$ *and satisfies the following two integral equations:*

$$\begin{cases} x_t = \pi + \int_0^t B(s, x_s, Y_s, z_s, Z_s, \xi_s, v_s)ds + \int_0^t \Sigma(s, x_s, Y_s, z_s, Z_s, \xi_s, v_s)dW_s \\ \qquad\qquad + \int_0^t \int_\Theta \Phi(s, x_s, Y_s, z_s, Z_s, \xi_s, v_s, \theta)\tilde{N}(d\theta, ds) - \int_0^t z_s \overleftarrow{dB}_s, \\ Y_t = h(x_T) + \int_t^T F(s, x_s, Y_s, z_s, Z_s, \xi_s, v_s)ds \\ \qquad\qquad + \int_t^T G(s, x_s, Y_s, z_s, Z_s, \xi_s, v_s)\overleftarrow{dB}_s \\ \qquad\qquad - \int_t^T Z_s dW_s - \int_t^T \int_\Theta \xi_s(\theta)\tilde{N}(d\theta, ds), \ t \in [0, T]. \end{cases}$$

In Sect. 4 we shall discuss the existence and uniqueness of (1).

Let us now introduce a *cost functional*:

$$J(v.) := \mathbb{E}\Big[\int_0^T \ell(t, x_t, Y_t, z_t, Z_t, \xi_t, v_t)dt + \varphi(x_T) + \psi(Y_0)\Big], \ v. \in \mathcal{U}_{ad}, \tag{2}$$

with

$$\varphi, \psi : H \to \mathbb{R},$$
$$\ell : \Omega \times [0, T] \times K \times K \times L_2(H_2; K) \times L_2(H_1; K) \times L^2_\nu(K) \times \mathcal{O} \to \mathbb{R},$$

being measurable functions such that (2) is defined. See assumption (A3) below for precise assumptions.

The control problem of system (1) is to minimize J over $\mathcal{U}_{ad}$. Thus an admissible control $u.$ is called an *optimal control* if

$$J(u.) = \inf_{v. \in \mathcal{U}_{ad}} J(v.). \tag{3}$$

In this case we shall say that $(x, Y, z, Z, \xi, u.)$ is an *optimal solution* of the control problem (1)–(3).

Further details on this control problem will be the main purpose of Sect. 5. We discuss next the existence and uniqueness of (1).

4 Forward-Backward Doubly Stochastic Differential Equations

We shall be interested here in the existence and uniqueness of the solution to FBDSDE (1). Keeping the notations in Sect. 3 denote

$$A(t, X, v) = (-F, B, -G, \Sigma, \Phi)(t, X, v)$$

and

$$\langle A, X\rangle = -\langle x, F\rangle_K + \langle Y, B\rangle_K - \langle z, G\rangle_{L_2(H_2;K)} + \langle Z, \Sigma\rangle_{L_2(H_1;K)} + \langle \xi, \Phi\rangle_{L^2_\nu(K)},$$

for $X = (x, Y, z, Z, \xi) \in K \times K \times L_2(H_2; K) \times L_2(H_1; K) \times L^2_\nu(K)$ and $(t, v) \in [0, T] \times \mathcal{O}$. The following three assumptions on the coefficients of system (1) and (2) are our main assumptions.

(A1) $\forall X = (x, Y, z, Z, \xi), \bar{X} = (\bar{x}, \bar{Y}, \bar{z}, \bar{Z}, \bar{\xi}) \in K \times K \times L_2(H_2; K) \times L_2(H_1; K) \times L^2_\nu(K)$, $\forall t \in [0, T]$, $\forall v \in \mathcal{O}$,

$$\begin{aligned}\left\langle A(t, X, v) - A(t, \bar{X}, v), X - \bar{X} \right\rangle \leq -\lambda(|x - \bar{x}|^2_K + \left|Y - \bar{Y}\right|^2_K \\ + \|z - \bar{z}\|^2_{L_2(H_2;K)} + \left\|Z - \bar{Z}\right\|^2_{L_2(H_1;K)} + \left\|\xi - \bar{\xi}\right\|^2_{L^2_\nu(K)}),\end{aligned}$$

and

$$c > 0,$$

or

(A1)'

$$\begin{aligned}\left\langle A(t, X, v) - A(t, \bar{X}, v), X - \bar{X} \right\rangle \geq \lambda(|x - \bar{x}|^2_K + \left|Y - \bar{Y}\right|^2_K \\ + \|z - \bar{z}\|^2_{L_2(H_2;K)} + \left\|Z - \bar{Z}\right\|^2_{L_2(H_1;K)} + \left\|\xi - \bar{\xi}\right\|^2_{L^2_\nu(K)}),\end{aligned}$$

and

$$c < 0,$$

for some $\lambda > 0$.

(A2) For each $X \in K \times K \times L_2(H_2; K) \times L_2(H_1; K) \times L^2_\nu(K)$ and $v \in \mathcal{O}$, we have $A(\cdot, X, v) \in \mathbb{M}^2$.

(A3) We have

$$\begin{cases}(i)\ F, B, G, \Sigma, \Phi, \ell \text{ are continuously differentiable with respect to } (x, Y, z, Z, \xi), \\ \quad \text{also } \varphi \text{ and } \psi \text{ are continuously differentiable with respect to } x \text{ and } Y, \\ \quad \text{respectively,} \\ (ii)\ \text{the derivatives of } F, B, G, \Sigma, \Phi \text{ with respect to the above arguments are} \\ \quad \text{bounded,} \\ (iii)\ \text{the derivatives of } \ell \text{ are bounded by } C(1 + |x| + |Y| + \|z\| + \|Z\| + |||\xi|||), \\ (iv)\ \varphi_x \text{ and } \psi_Y \text{ are bounded by } C\,(1 + |x|) \text{ and } C\,(1 + |Y|), \text{ respectively,}\end{cases}$$

for some constant $C > 0$.

Remark 1. The condition $c > 0$ in (A1) guarantees the following monotonicity condition of the mapping h:

$$\langle h(x) - h(\bar{x}), x - \bar{x} \rangle_K \geq c\,|x - \bar{x}|^2_K, \quad \forall\, x, \bar{x} \in K.$$

A similar thing happens also for the case $c < 0$ in (A1)'.

Theorem 2. *If (A1)–(A3) (or (A1)', (A2)–(A3)) hold, then there exists a unique solution (x, Y, z, Z, ξ) of the FBDSDE (1).*

By making use of the extended martingale representation theorem (Theorem 1), the proof is standard and can be achieved directly by following the outline of the proofs in [13,14]. So we omit it.

This theorem in its infinite dimensional setting is new. In fact, as far as know, this is the first appearance of such an infinite dimensional result as well as Theorem 1.

5 Stochastic Maximum Principle

To derive the maximum principle we define the *Hamiltonian* $\mathcal{H}$ from $[0,T] \times \Omega \times K \times K \times L_2(H_2;K) \times L_2(H_1;K) \times L^2_\nu(K) \times \mathcal{O} \times K \times K \times L_2(H_2;K) \times L_2(H_1;K) \times L^2_\nu(K)$ to $\mathbb{R}$ by the formula:

$$\begin{aligned}\mathcal{H}(t,x,Y,z,Z,\xi,v,p,P,q,Q,\Upsilon) &:= -\langle p, F(t,x,Y,z,Z,\xi,v)\rangle_K \\ &+ \langle P, B(t,x,Y,z,Z,\xi,v)\rangle_K - \langle q, G(t,x,Y,z,Z,\xi,v)\rangle_{L_2(H_2;K)} \\ &+ \langle Q, \Sigma(t,x,Y,z,Z,\xi,v)\rangle_{L_2(H_1;K)} + \ell(t,x,Y,z,Z,\xi,v) \\ &+ \int_\Theta \left\langle \Upsilon(\widehat{\theta}), \Phi(t,x,Y,z,Z,\xi,v,\widehat{\theta}) \right\rangle_{L^2_\nu(K)} \nu(d\widehat{\theta}). \end{aligned} \tag{4}$$

Theorem 3. *Let $v.$ be an arbitrary element of $\mathcal{U}_{ad}$. Assume (A1)–(A3). Let $\{(y_t, Y_t, z_t, Z_t, k_t),\ t \in [0,T]\}$ be the corresponding solution of (1). Then there exists a unique solution (p,P,q,Q,Υ) of the following adjoint equations of (1):*

$$\begin{cases} dp_t = -\mathcal{H}_Y(t)dt - \mathcal{H}_Z dW_t - q_t \overleftarrow{dB}_t - \int_\Theta \mathcal{H}_\xi(t)\tilde{N}(d\theta,dt), \\ dP_t = -\mathcal{H}_x(t)dt - \mathcal{H}_z(t)\overleftarrow{dB}_t + Q_t dW_t + \int_\Theta \Upsilon_t(\theta)\tilde{N}(d\theta,dt), \\ p_0 = -\psi_Y(Y_0), P_T = -\,c\,p_T + \varphi_x(x_T), \end{cases} \tag{5}$$

where $\mathcal{H}_x(t)$ is the gradient $\nabla_x \mathcal{H}(t,x,Y_t,z_t,Z_t,\xi_t,v_t,p_t,P_t,q_t,Q_t,\Upsilon_t) \in K, \ldots$ etc.

Proof. Thanks to assumptions (A1)–(A3) this linear FBDSDE satisfy (A1)', (A2) and (A3). In fact the monotonicity condition follows from the definition of Gâteaux derivatives (as limits) and the fact that the corresponding mappings satisfy originally the monotonicity condition in (A1). Hence the result follows from Theorem 2.

We are now ready to state the stochastic maximum principle for the optimal control problem (1)–(3).

Theorem 4. *Suppose that (A1)–(A3) hold. Given $u. \in \mathcal{U}_{ad}$, let $(x^{u.}, Y^{u.}, z^{u.}, Z^{u.}, \xi^{u.})$ and $(p^{u.}, P^{u.}, q^{u.}, Q^{u.}, \Upsilon^{u.})$ be the corresponding solutions of FBDSDEs (1) and (5), respectively. Assume that the following assumptions hold.*
(i) φ and ψ are convex;
(ii) For all $t \in [0,T]$, $\mathbb{P}$ - a.s., the function $\mathcal{H}(t,\cdot,\cdot,\cdot,\cdot,\cdot,\cdot,p^{u.},P^{u.},q^{u.},Q^{u.},\Upsilon^{u.})$ is convex;
(iii) We have

$$\begin{aligned}&\mathcal{H}(t,x_t^{u.},Y_t^{u.},z_t^{u.},Z_t^{u.},\xi_t^{u.},u_t,p_t^{u.},P_t^{u.},q_t^{u.},Q_t^{u.},\Upsilon_t^{u.}) \\ &\quad = \inf_{v\in U} \mathcal{H}(t,x_t^{u.},Y_t^{u.},z_t^{u.},Z_t^{u.},\xi_t^{u.},v,p_t^{u.},P_t^{u.},q_t^{u.},Q_t^{u.},\Upsilon_t^{u.}),\end{aligned} \tag{6}$$

for a.e. t, $\mathbb{P}$ - a.s.

Then $(x^{u.}, Y^{u.}, z^{u.}, Z^{u.}, \xi^{u.}, u.)$ is an optimal solution of the control problem (1)–(3).

Proof. Let $v.$ be an arbitrary element of $\mathcal{U}_{ad}$. With the help of assumptions (A1)–(A3) let, by using of Theorem 2, $(x^{v.}, Y^{v.}, z^{v.}, Z^{v.}, \xi^{v.})$ be the corresponding solution of FBDSDE (1). Applying (2), the convexity of φ and ψ, the adjoint equations (5) and system (1) it follows that

$$J(v.) - J(u.) \geq \mathbb{E}\big[\,\langle P_T^{u.} + c\, p_T^{u.}, x_T^{v.} - x_T^{u.}\rangle\,\big] - \mathbb{E}\big[\,\langle p_0^{u.}, Y_0^{v.} - Y_0^{u.}\rangle\,\big]$$
$$+\mathbb{E}\Big[\int_0^T \big(\ell(t, x_t^{v.}, Y_t^{v.}, z_t^{v.}, Z_t^{v.}, \xi_t^{v.}, v_t) - \ell(t, x_t^{u.}, Y_t^{u.}, z_t^{u.}, Z_t^{u.}, \xi_t^{u.}, u_t)\big)dt\Big]. \quad (7)$$

Next, by applying a suitable Itô's formula for infinite dimensional SDEs driven by Wiener processes on Hilbert spaces and Poisson measures to compute $\langle p_t^{u.}, Y_t^{v.} - Y_t^{u.}\rangle_K$ and $\langle P_t^{u.}, x_t^{v.} - x_t^{u.}\rangle_K$, we derive with the help of assumptions (A2) and (A3) that

$$\mathbb{E}\big[\,\langle P_T^{u.}, x^{v.} - x_T^{u.}\rangle\,\big] - \mathbb{E}\big[\,\langle p_0^{u.}, Y_0^{v.} - Y_0^{u.}\rangle\,\big] = -\mathbb{E}\big[\,\langle p_T^{u.}, Y_T^{v.} - Y_T^{u.}\rangle\,\big]$$
$$- \mathbb{E}\Big[\int_0^T \langle p_t^{u.}, F(t, x_t^{v.}, Y_t^{v.}, z_t^{v.}, Z_t^{v.}, \xi_t^{v.}, v_t) - F(t, x_t^{u.}, Y_t^{u.}, z_t^{u.}, Z_t^{u.}, \xi_t^{u.}, u_t)\rangle\, dt\Big]$$
$$+ \mathbb{E}\Big[\int_0^T \langle \mathcal{H}_Y(t, x_t^{u.}, Y_t^{u.}, z_t^{u.}, Z_t^{u.}, \xi_t^{u.}, u_t, p_t^{u.}, P_t^{u.}, q_t^{u.}, Q_t^{u.}, \Upsilon_t^{u.}), Y_t^{v.} - Y_t^{u.}\rangle\, dt\Big]$$
$$- \mathbb{E}\Big[\int_0^T \langle q_t^{u.}, G(t, x_t^{v.}, Y_t^{v.}, z_t^{v.}, Z_t^{v.}, \xi_t^{v.}, v_t) - G(t, x_t^{u.}, Y_t^{u.}, z_t^{u.}, Z_t^{u.}, \xi_t^{u.}, u_t)\rangle\, dt\Big]$$
$$+ \mathbb{E}\Big[\int_0^T \langle \mathcal{H}_Z(t, x_t^{u.}, Y_t^{u.}, z_t^{u.}, Z_t^{u.}, \xi_t^{u.}, u_t, p_t^{u.}, P_t^{u.}, q_t^{u.}, Q_t^{u.}, \Upsilon_t^{u.}), Z_t^{v.} - Z_t^{u.}\rangle\, dt\Big]$$
$$+ \mathbb{E}\Big[\int_0^T \int_\Theta \langle \mathcal{H}_\xi(t, x_t^{u.}, Y_t^{u.}, z_t^{u.}, Z_t^{u.}, \xi_t^{u.}, u_t, p_t^{u.}, P_t^{u.}, q_t^{u.}, Q_t^{u.}, \Upsilon_t^{u.}),$$
$$\xi_t^{v.}(\theta) - \xi_t^{u.}(\theta)\rangle \nu(d\theta)dt\Big]$$
$$+ \mathbb{E}\Big[\int_0^T \langle P_t^{u.}, B(t, x_t^{v.}, Y_t^{v.}, z_t^{v.}, Z_t^{v.}, \xi_t^{v.}, v_t)$$
$$- B(t, x_t^{u.}, Y_t^{u.}, z_t^{u.}, Z_t^{u.}, \xi_t^{u.}, u_t)\rangle dt\Big]$$
$$+ \mathbb{E}\Big[\int_0^T \langle \mathcal{H}_x(t, x_t^{u.}, Y_t^{u.}, z_t^{u.}, Z_t^{u.}, \xi_t^{u.}, u_t, p_t^{u.}, P_t^{u.}, q_t^{u.}, Q_t^{u.}, \Upsilon_t^{u.}), x_t^{v.} - x_t^{u.}\rangle\, dt\Big]$$
$$+ \mathbb{E}\Big[\int_0^T \langle \mathcal{H}_z(t, x_t^{u.}, Y_t^{u.}, z_t^{u.}, Z_t^{u.}, \xi_t^{u.}, u_t, p_t^{u.}, P_t^{u.}, q_t^{u.}, Q_t^{u.}, \Upsilon_t^{u.}), z_t^{v.} - z_t^{u.}\rangle\, dt\Big]$$
$$+ \mathbb{E}\Big[\int_0^T \langle Q_t^{u.}, \Sigma(t, x_t^{v.}, Y_t^{v.}, z_t^{v.}, Z_t^{v.}, \xi_t^{v.}, v_t) - \Sigma(t, x_t^{u.}, Y_t^{u.}, z_t^{u.}, Z_t^{u.}, \xi_t^{u.}, u_t)\rangle\, dt\Big]$$
$$+ \mathbb{E}\Big[\int_0^T \int_\Theta \langle \Upsilon_t^{u.}(\theta), \Phi(t, x_t^{v.}, Y_t^{v.}, z_t^{v.}, Z_t^{v.}, \xi_t^{v.}, v_t, \theta)$$
$$- \Phi(t, x_t^{u.}, Y_t^{u.}, z_t^{u.}, Z_t^{u.}, \xi_t^{u.}, u_t, \theta)\rangle \nu(d\theta)dt\Big]. \quad (8)$$

On the hand, from the formula $h(\omega, x) := c\,x + \xi(\omega), x \in K$, one gets easily the cancelation:

$$\mathbb{E}\big[\,\langle c\,p_T^{u.}, x_T^{v.} - x_T^{u.}\rangle\,\big] - \mathbb{E}\big[\,\langle p_T^{u.}, Y_T^{v.} - Y_T^{u.}\rangle\,\big] = 0. \tag{9}$$

Therefore, by applying (8) and (9) in (7), using the formula of $\mathcal{H}$ in (4) and then the convexity of $\mathcal{H}$ in condition (ii) we obtain

$$\begin{aligned} &J(v.) - J(u.) \\ &\geq \mathbb{E}\big[\int_0^T \langle \mathcal{H}_v(t, x_t^{u.}, Y_t^{u.}, z_t^{u.}, Z_t^{u.}, \xi_t^{u.}, u_t, p_t^{u.}, P_t^{u.}, q_t^{u.}, Q_t^{u.}, \Upsilon_t^{u.}), v_t - u_t\rangle_{\mathcal{O}}\, dt\big]. \end{aligned} \tag{10}$$

But the minimum condition (iii) yields

$$\langle \mathcal{H}_v(t, x_t^{u.}, Y_t^{u.}, z_t^{u.}, Z_t^{u.}, \xi_t^{u.}, u_t, p_t^{u.}, P_t^{u.}, q_t^{u.}, Q_t^{u.}, \Upsilon_t^{u.}), v_t - u_t\rangle_{\mathcal{O}} \geq 0.$$

Consequently (10) becomes

$$J(v.) - J(u.) \geq 0.$$

Since $v.$ is an arbitrary element of $\mathcal{U}_{ad}$, then $u.$ is an optimal control, and so the proof is complete.

Remark 2. Condition (A1) assumed in Theorem 4 is only needed to guarantee the existence and uniqueness of the solutions of (1) and (5), and so if one can get such solutions without assuming (A1) there will not any necessity to assume it in advance in this theorem.

References

1. Ahmed, N.U., Charalambous, C.D.: Stochastic minimum principle for partially observed systems subject to continuous and jump diffusion processes and driven by relaxed controls. arXiv:1302.3455v1 [math.OC] (2013)
2. Al-Hussein, A.: Martingale representation theorem in infinite dimensions. Arab. J. Math. Sci. **10**(1), 1–18 (2004)
3. Al-Hussein, A.: Backward stochastic differential equations in infinite dimensions and applications. Arab. J. Math. Sci. **10**(2), 1–42 (2004)
4. Al-Hussein, A.: Sufficient conditions of optimality for backward stochastic evolution equations. Commun. Stoch. Anal. **4**(3), 433–443 (2010)
5. Al-Hussein, A., Gherbal, B.: Maximum principle for optimal control of forward-backward doubly stochastic differential equations with jumps. arXiv:1301.1948v4 [math.OC] (2013, submitted)
6. Hu, Y., Peng, S.: Maximum principle for semilinear stochastic evolution equation control systems. Stochastics **33**, 159–180 (1990)
7. Hu, Y., Peng, S.: Solutions of forward-backward stochastic differential equations. Probab. Theor. Relat. Fields **103**, 273–283 (1995)
8. Ikeda, N., Watanabe, S.: Stochastic Differential Equations and Diffusion Processes. North Holland/Kodansha, Amsterdam/Tokyo (1981)

9. Ma, J., Yong, J.: Forward-backward Stochastic Differential Equations and Their Applications. Lecture Notes in Mathematics, vol. 1702. Springer, Berlin (1999)
10. Øksendal, B., Proske, F., Zhang, T.: Backward stochastic partial differential equations with jumps and application to optimal control of random jump fields. Stochastics **77**(5), 381–399 (2005)
11. Pardoux, E., Peng, S.: Backward doubly stochastic differential equations and systems of quasilinear SPDEs. Probab. Theor. Relat. Fields **98**, 209–227 (1994)
12. Peng, S., Shi, Y.: A type-symmetric forward-backward stochastic differential equations. C. R. Acad. Sci. Paris Ser. I **336**(1), 773–778 (2003)
13. Peng, S., Wu, Z.: Fully coupled forward-backward stochastic differential equations and applications to optimal control. SIAM J. Control Optim. **37**, 825–843 (1999)
14. Zhu, Q., Shi, Y.: Forward-backward doubly stochastic differential equations with random jumps and stochastic partial differential-integral equations, [OL].[201001-1044]. http://www.paper.edu.cn/index.php/default/en-releasepaper/downPaper/201001-1044
15. Shi, J.T., Wu, Z.: Maximum principle for partially-observed optimal control of fully-coupled forward-backward stochastic systems. J Optim. Theor. Appl. **145**, 543–578 (2010)
16. Situ, R.: On solutions of backward stochastic differential equations with jumps and with non-Lipschitzian coefficients in Hilbert spaces and stochastic control. Stat. Probab. Lett. **60**(4), 279–288 (2002)
17. Tessitore, G.: Existence, uniqueness and space regularity of the adapted solutions of a backward SPDE. Stoch. Anal. Appl. **14**(4), 461–486 (1996)
18. Wang, X., Wu, Z.: FBSDE with Poisson process and its application to linear quadratic stochastic optimal control problem with random jumps. Acta Automatica Sinica **29**, 821–826 (2003)
19. Yin, J., Wang, Y.: Hilbert space-valued forward-backward stochastic differential equations with Poisson jumps and applications. Math. Anal. Appl. **328**, 438–451 (2007)

Efficient Solvers for Large-Scale Saddle Point Systems Arising in Feedback Stabilization of Multi-field Flow Problems

Peter Benner(✉), Jens Saak, Martin Stoll, and Heiko K. Weichelt

Max Planck Institute for Dynamics of Complex Technical Systems,
Sandtorstr. 1, 39106 Magdeburg, Germany
{benner,saak,stoll,weichelt}@mpi-magdeburg.mpg.de
http://www.mpi-magdeburg.mpg.de

Abstract. This article introduces a block preconditioner to solve large-scale block structured saddle point systems using a Krylov-based method. Such saddle point systems arise, e.g., in the Riccati-based feedback stabilization approach for multi-field flow problems as discussed in [2]. Combining well known approximation methods like a least-squares commutator approach for the Navier-Stokes Schur complement, an algebraic multigrid method, and a Chebyshev-Semi-Iteration, an efficient preconditioner is derived and tested for different parameter sets by using a simplified reactor model that describes the spread concentration of a reactive species forced by an incompressible velocity field.

Keywords: Coupled flow control · Large-scale saddle point systems · Preconditioned GMRES · Least-squares commutator approach · Algebraic multigrid · Chebyshev-Semi-Iteration

1 Introduction

In this paper we investigate the solution of large-scale saddle point systems arising in control problems for coupled partial differential equations (PDEs). The starting points are recent publications concerning the boundary feedback stabilization of non-coupled flows like the linear Stokes flow in [3] and the non-linear Navier-Stokes flow in [1]. The analytic approach for this feedback stabilization is given by Raymond in, e.g., [13].

Using the projection idea proposed by Heinkenschloss et al. [8], Benner et al. [1,3] show that the solution of certain saddle point systems is the key ingredient to ensure that the numerical solution lies on the correct solution manifold, i.e., the space of discretely divergence-free velocity fields, without performing an explicit projection.

Applying these ideas to a coupled flow problem, namely the Navier-Stokes equations combined with a diffusion-convection equation, leads to saddle point systems with a more complicated block structure [2]. Solving these systems efficiently requires the use of appropriate preconditioners. This paper investigates

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 11–20, 2014.
DOI: 10.1007/978-3-662-45504-3_2

an efficient iterative solution strategy via the use of preconditioned Krylov subspace methods based on the framework derived in [3]. Here we consider the full feedback system for the coupled multi-field flow problem, while in [3], only the linear Stokes case was treated without coupling to another field equation. Moreover, this paper complements [2] in the sense that there, we have focused on presenting results on the convergence of the Newton-ADI method for solving the algebraic Riccati equation determining the stabilizing feedback control for the coupled system, where the saddle point problems in the innermost step of the Newton-ADI iteration were solved by sparse direct methods, while here, we study preconditioned iterative solvers for this step.

The remainder of this paper is organized as follows. Section 2 briefly recalls the feedback stabilization approach for multi-field flow problems from [2] that leads to large-scale saddle point systems. Afterwards, we discuss properties of these saddle point systems to derive a suitable preconditioner in Sect. 3. Section 4 shows numerical results before we conclude the paper and give a short outlook to further investigations in Sect. 5.

2 Derivation of Saddle Point Systems

The derivation of the block structured saddle point systems in [2] starts with the linearized coupled flow problem defined for $t \in [0, \infty)$ and $\boldsymbol{x} \in \Omega \subset \mathbb{R}^2$. The linearized Navier-Stokes equations that describe, up to first order, the difference between actual and desired velocity and pressure are given as

$$\begin{aligned} \frac{\partial}{\partial t}\boldsymbol{z} - \frac{1}{\mathrm{Re}}\Delta\boldsymbol{z} + (\boldsymbol{w}\cdot\nabla)\boldsymbol{z} + (\boldsymbol{z}\cdot\nabla)\boldsymbol{w} + \nabla p &= \boldsymbol{f}_l, \\ \operatorname{div}\boldsymbol{z} &= 0, \end{aligned} \qquad \text{on } [0,\infty)\times\Omega. \tag{1}$$

They are then coupled via the velocity field $\boldsymbol{z}(t, \boldsymbol{x})$ with the linearized diffusion-convection equation

$$\frac{\partial}{\partial t}c_{\boldsymbol{z}} - \frac{1}{\mathrm{ReSc}}\Delta c_{\boldsymbol{z}} + (\boldsymbol{w}\cdot\nabla)c_{\boldsymbol{z}} + (\boldsymbol{z}\cdot\nabla)c_{\boldsymbol{w}} = 0, \quad \text{on } [0,\infty)\times\Omega \tag{2}$$

that describes the concentration of a reactive species denoted by $c_{\mathbf{z}}(t, \boldsymbol{x})$. The stationary linearization points $\boldsymbol{w}(\boldsymbol{x})$ for the velocity and $c_{\boldsymbol{w}}(\boldsymbol{x})$ for the concentration are assumed to be given. The equations are scaled with the Reynolds number Re and the Schmidt number Sc. Using the mixed Taylor-Hood finite elements [9] for the velocity and pressure in Eq. (1) as well as linear ansatz functions for the concentration in Eq. (2), we end up with a system of discrete differential-algebraic equations (DAE) that can be written as the control system:

$$\begin{bmatrix} M_{\mathbf{z}} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & M_{\mathbf{c}} \end{bmatrix} \frac{d}{dt}\begin{bmatrix} \mathbf{z} \\ \mathbf{p} \\ \mathbf{c} \end{bmatrix} = \begin{bmatrix} A_{\mathbf{z}} & G & 0 \\ G^T & 0 & 0 \\ -R & 0 & A_{\mathbf{c}} \end{bmatrix}\begin{bmatrix} \mathbf{z} \\ \mathbf{p} \\ \mathbf{c} \end{bmatrix} + \begin{bmatrix} B_{\mathbf{z}} \\ 0 \\ 0 \end{bmatrix}\mathbf{u}, \tag{3a}$$

$$\mathbf{y} = \begin{bmatrix} 0 & 0 & C_{\mathbf{c}} \end{bmatrix}\begin{bmatrix} \mathbf{z} \\ \mathbf{p} \\ \mathbf{c} \end{bmatrix} \tag{3b}$$

with the first block row for velocity (of dimension $n_{\mathbf{z}}$), the second row for pressure (of dimension $n_{\mathbf{p}}$), and the third row for concentration (of dimension $n_{\mathbf{c}}$) [2].

The matrix pencil

$$\left(\underbrace{\begin{bmatrix} A_{\mathbf{z}} & G & 0 \\ G^T & 0 & 0 \\ -R & 0 & A_{\mathbf{c}} \end{bmatrix}}_{\mathbf{A}}, \underbrace{\begin{bmatrix} M_{\mathbf{z}} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & M_{\mathbf{c}} \end{bmatrix}}_{\mathbf{M}} \right)$$

is of dimension $n \times n$ with $n = n_{\mathbf{z}} + n_{\mathbf{c}} + n_{\mathbf{p}}$ and has $2n_{\mathbf{p}}$ infinite eigenvalues [5].

In [2], a linear-quadratic regulator (LQR) approach is applied to system (3) for determining the stabilizing control function $\mathbf{u}$. The solution of this LQR problem is a linear feedback control $\mathbf{u}(t) = \mathbf{K}(\mathbf{z}(t), \mathbf{p}(t), \mathbf{c}(t))$, determined via the solution of an algebraic Riccati equation (ARE) defined on the subspace of discretely divergence-free vector fields. The resulting ARE is then solved using a Newton-ADI algorithm. This method yields a threefold nested iteration. In the innermost loop, saddle point systems of the form

$$\underbrace{\begin{bmatrix} A_{\mathbf{z}}^T + q_i M_{\mathbf{z}} & G & -R^T \\ G^T & 0 & 0 \\ 0 & 0 & A_{\mathbf{c}}^T + q_i M_{\mathbf{c}} \end{bmatrix}}_{=\mathbf{A}^T + q_i \mathbf{M} =: \mathbf{F}_i} \underbrace{\begin{bmatrix} \Lambda_{\mathbf{z}} \\ \Lambda_{\mathbf{p}} \\ \Lambda_{\mathbf{c}} \end{bmatrix}}_{\mathbf{\Lambda}} = \underbrace{\begin{bmatrix} \tilde{Y}_{\mathbf{z}} \\ 0 \\ \tilde{Y}_{\mathbf{c}} \end{bmatrix}}_{\mathbf{Y}}. \tag{4}$$

have to be solved for certain ADI shifts $q_i \in \mathbb{C}^-$ and a block right hand side $\mathbf{Y}$. The whole nested iteration is given in [2, Algorithm 1] and is omitted here due to space constraints.

3 Preconditioned Iterative Solvers for Block Structured Saddle Point Systems

The use of direct solvers in (4) is only suitable for moderate problem sizes and two-dimensional problems. Although iterative methods can handle much larger systems, their performance will deteriorate if the mesh-size decreases. To avoid this, a suitable preconditioner $\mathbf{P}_i \in \mathbb{C}^{n \times n}$ is introduced such that

$$\mathbf{P}_i^{-1} \mathbf{F}_i \mathbf{\Lambda} = \mathbf{P}_i^{-1} \mathbf{Y}$$

is solved instead of (4) (see [7,16]). Before we derive a suitable preconditioner $\mathbf{P}_i$ we need to describe the properties of the saddle point system and their influence on the chosen preconditioner.

3.1 Properties

The matrices $M_{\mathbf{z}}, M_{\mathbf{c}}$ are symmetric and positive definite, G, R are of full rank, and the ADI shift $q_i \in \mathbb{C}^-$ is contained in the convex hull of the finite spectrum of $(\mathbf{A}, \mathbf{M})$. The shifted system matrix $\mathbf{F}_i$ is indefinite $\forall q_i \in \mathbb{C}^-$. Due to the different q_i, the matrix $\mathbf{F}_i$ changes in each ADI step and, therefore, the preconditioner has to be adapted in each ADI step as well. Nevertheless, for the remainder of this section we assume a fixed ADI shift $q_i = q$ to omit the index i if it is obvious.

3.2 Derivation of Block Preconditioner

Adapting the ideas from [3, Sect. 3.2] we consider

$$\mathbf{F} = \begin{bmatrix} F_{\mathbf{z}} & G & -R^T \\ G^T & 0 & 0 \\ 0 & 0 & F_{\mathbf{c}} \end{bmatrix} = \begin{bmatrix} \mathbf{F}_{NSE} & -\tilde{R}^T \\ 0 & F_{\mathbf{c}} \end{bmatrix} \quad \text{with} \quad \begin{aligned} F_{\mathbf{z}} &:= A_{\mathbf{z}}^T + qM_{\mathbf{z}}, \\ F_{\mathbf{c}} &:= A_{\mathbf{c}}^T + qM_{\mathbf{c}}, \\ \tilde{R} &:= \begin{bmatrix} R & 0 \end{bmatrix}, \end{aligned} \tag{5}$$

and $\mathbf{F}_{NSE}$ as the saddle point system for the non-coupled Navier-Stokes flow as it is used in [1]. Using the preconditioner $\mathbf{P}_{NSE}$ from [3], we define a block preconditioner for the use with GMRES [17] to solve with the block structured saddle point system (5) as follows:

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{NSE} & -\tilde{R}^T \\ 0 & P_{\mathbf{c}} \end{bmatrix} = \begin{bmatrix} P_{\mathbf{z}} & 0 & -R^T \\ G^T & -P_{SC} & 0 \\ 0 & 0 & P_{\mathbf{c}} \end{bmatrix}$$

$$\Rightarrow \mathbf{P}^{-1} = \begin{bmatrix} P_{\mathbf{z}}^{-1} & 0 & P_{\mathbf{z}}^{-1}R^T P_{\mathbf{c}}^{-1} \\ P_{SC}^{-1}G^T P_{\mathbf{z}}^{-1} & -P_{SC}^{-1} & P_{SC}^{-1}G^T P_{\mathbf{z}}^{-1}R^T P_{\mathbf{c}}^{-1} \\ 0 & 0 & P_{\mathbf{c}}^{-1} \end{bmatrix}.$$

In contrast to the preconditioner derived in [3], we cannot achieve a block lower triangular matrix due to the coupling matrix R. Applying $\mathbf{P}^{-1}$ to $\mathbf{F}$ yields

$$\mathbf{P}^{-1}\mathbf{F} = \begin{bmatrix} P_{\mathbf{z}}^{-1}F_{\mathbf{z}} & P_{\mathbf{z}}^{-1}G & -P_{\mathbf{z}}^{-1}R^T + P_{\mathbf{z}}^{-1}R^T P_{\mathbf{c}}^{-1}F_{\mathbf{c}} \\ P_{SC}^{-1}G^T P_{\mathbf{z}}^{-1}F_{\mathbf{z}} - P_{SC}^{-1}G^T & P_{SC}^{-1}G^T P_{\mathbf{z}}^{-1}G & -P_{SC}^{-1}G^T P_{\mathbf{z}}^{-1}R^T + P_{SC}^{-1}G^T P_{\mathbf{z}}^{-1}R^T P_{\mathbf{c}}^{-1}F_{\mathbf{c}} \\ 0 & 0 & P_{\mathbf{c}}^{-1}F_{\mathbf{c}} \end{bmatrix} \tag{6}$$

If one assumes $P_{\mathbf{z}} = F_{\mathbf{z}}$, $P_{\mathbf{c}} = F_{\mathbf{c}}$, and $P_{SC} = G^T F_{\mathbf{z}}^{-1} G$ as ideal approximations in (6), this leads to

$$\begin{bmatrix} I_{\mathbf{z}} & F_{\mathbf{z}}^{-1}G & -F_{\mathbf{z}}^{-1}R^T + F_{\mathbf{z}}^{-1}R^T \\ P_{SC}^{-1}G^T - P_{SC}^{-1}G^T & P_{SC}^{-1}G^T F_{\mathbf{z}}^{-1}G & -P_{SC}^{-1}G^T F_{\mathbf{z}}^{-1}R^T + P_{SC}^{-1}G^T F_{\mathbf{z}}^{-1}R^T \\ 0 & 0 & I_{\mathbf{c}} \end{bmatrix}$$

$$= \begin{bmatrix} I_{\mathbf{z}} & * & 0 \\ 0 & I_{\mathbf{p}} & 0 \\ 0 & 0 & I_{\mathbf{c}} \end{bmatrix}$$

and our iterative method would converge within one step. The goal is to find good approximations for $P_{\mathbf{z}}, P_{\mathbf{c}}$, and P_{SC} that can be evaluated fast and still cluster the eigenvalues in a suitable way such that our iterative solver shows fast convergence [7]. Instead of calculating the inverse $\mathbf{P}^{-1}$ to apply the preconditioner $\mathbf{P}$, we consider the solution of a linear system

$$\begin{bmatrix} P_{\mathbf{z}} & 0 & -R^T \\ G^T & -P_{SC} & 0 \\ 0 & 0 & P_{\mathbf{c}} \end{bmatrix} \begin{bmatrix} x_{\mathbf{z}} \\ x_{\mathbf{p}} \\ x_{\mathbf{c}} \end{bmatrix} = \begin{bmatrix} b_{\mathbf{z}} \\ b_{\mathbf{p}} \\ b_{\mathbf{c}} \end{bmatrix} \tag{7}$$

that can be solved in three steps:

$$\text{Step I:} \qquad x_{\mathbf{c}} = P_{\mathbf{c}}^{-1} b_{\mathbf{c}}, \tag{8a}$$

$$\text{Step II:} \qquad x_{\mathbf{z}} = P_{\mathbf{z}}^{-1}(R^T x_{\mathbf{c}} + b_{\mathbf{z}}), \tag{8b}$$

$$\text{Step III:} \qquad x_{\mathbf{p}} = P_{SC}^{-1}(G^T x_{\mathbf{z}} - b_{\mathbf{p}}). \tag{8c}$$

In conclusion, the coupling matrix R only leads to a matrix-vector multiplication. In steps I and II, one needs to solve with the shifted velocity and concentration system matrices as defined in (5). For both steps, an algebraic multigrid (AMG) method can be used as it is described below. But first, we discuss the more challenging step III that is handled as follows.

3.3 Approximation Methods

Schur Complement Approximation. P_{SC} is an approximation of the Navier-Stokes *Schur complement* $SC := G^T F_{\mathbf{z}}^{-1} G \in \mathbb{R}^{n_{\mathbf{p}} \times n_{\mathbf{p}}}$. Unfortunately, the matrix SC would be a dense matrix that includes the inverse of $F_{\mathbf{z}}$. To avoid the use of this matrix, we follow the approach in [3,18] and use a slightly modified variant of the least squares commutator approach as it is described in [7, Sect. 8.2]. Namely, we consider the shifted Oseen operator in the velocity space

$$\mathcal{F}_{\boldsymbol{z}} = -\frac{1}{\mathrm{Re}}\nabla^2 + \boldsymbol{w} \cdot \nabla + q\mathcal{I}.$$

Note that it is common practice to omit the reaction term $(\boldsymbol{z} \cdot \nabla)\boldsymbol{w}$ that appears in the linearized Navier-Stokes equations to derive preconditioners [7, Sect. 8]. Similar to [7, Sect. 8.2] and [6], we suppose that there exists an analogous operator on the pressure space defined as

$$\mathcal{F}_p = (-\frac{1}{\mathrm{Re}}\nabla^2 + \boldsymbol{w} \cdot \nabla + q\mathcal{I})_p.$$

The least squares commutator of the shifted Oseen operator with the gradient operator is defined as

$$\mathcal{E} = (\mathcal{F})\nabla - \nabla(\mathcal{F}_p)$$

and is supposed to become small in some sense [7]. Using the discrete versions of the operators, we end up with

$$E = (M_{\mathbf{z}}^{-1} F_{\mathbf{z}}) M_{\mathbf{z}}^{-1} G - M_{\mathbf{z}}^{-1} G (M_{\mathbf{p}}^{-1} F_{\mathbf{p}})$$

with $M_{\mathbf{p}}$ the mass matrix and $F_{\mathbf{p}} = A_{\mathbf{p}}^T + qM_{\mathbf{p}}$ the shifted system matrix, both defined on the pressure space. Premultiplying this by $G^T F_{\mathbf{z}}^{-1} M_{\mathbf{z}}$ and postmultiplying by $F_{\mathbf{p}}^{-1} M_{\mathbf{p}}$ yields [3]

$$G^T M_{\mathbf{z}}^{-1} G F_{\mathbf{p}}^{-1} M_{\mathbf{p}} \approx G^T F_{\mathbf{z}}^{-1} G = SC.$$

The large and dense matrix $G^T M_{\mathbf{z}}^{-1} G$ cannot be used explicitly, but it is shown in [7, Sect. 5.5.1] that this matrix is spectrally equivalent to the Laplacian $S_{\mathbf{p}}$ defined on the pressure space for an inf-sup stable discretization and an inflow-outflow problem [7, Sect. 8.2], as it is considered in this paper. Finally, we obtain

$$P_{SC} \approx S_{\mathbf{p}} F_{\mathbf{p}}^{-1} M_{\mathbf{p}} \quad \Rightarrow \quad P_{SC}^{-1} \approx M_{\mathbf{p}}^{-1} F_{\mathbf{p}} S_{\mathbf{p}}^{-1}.$$

In [4] the authors use a similar approach for the Navier-Stokes equations. In summary, the application of P_{SC}^{-1} requires to solve with $S_{\mathbf{p}}$ (step IIIa), multiply with $F_{\mathbf{p}}$ (step IIIb), and solve with $M_{\mathbf{p}}$ (step IIIc). The step IIIa can be solved with an AMG method, similar to the steps I and II.

Algebraic Multigrid. As it is described above, the steps I (8a), II (8b) and IIIa are solved using an AMG method [14]. Due to the possibly complex ADI shifts q in (8a) and (8b), we use the AGMG package developed by the group of Y. Notay [10–12]. In all three cases we use the MATLAB®-based implementation to solve systems of the form

$$Fx = b$$

with a sparse matrix $F \in \{F_{\mathbf{z}}, F_{\mathbf{c}}, S_{\mathbf{p}}\}$. Details about the used parameters for the function `agmg` are discussed in Subsect. 4.2. For more details about the internally used methods and the implemented syntax we refer the reader to [11]. Although the AGMG method can handle complex arithmetic, it needs significantly more steps to converge to the desired tolerance. Additionally, we note that `agmg` is a non-linear function, such that one should use a flexible iterative method, e.g., FGMRES [15]. However, our numerical experiments do not show any drawbacks using a standard GMRES implementation.

Chebyshev-Semi-Iteration. Although the solution of step IIIc with the symmetric positive definite mass matrix $M_{\mathbf{p}}$ is relatively cheap, this can still be accelerated by using the *Chebyshev-Semi-Iteration* as it is described, e.g., in [18]. Numerical tests showed that one needs only 4−6 steps to obtain a suitable result for the preconditioner, which results in a speedup that is shown in Subsect. 4.2.

The next section depicts selected results to show the performance of the preconditioned iterative method.

4 Numerical Examples

To test the efficiency of the preconditioned iterative method, the same data and configurations as in [2] are used. After refining the initial triangulation of the reactor model in Fig. 1, we end up with the variable dimensions as depicted in Table 1b. Furthermore, we define five parameter sets for different combinations of Reynolds and Schmidt numbers as shown in Table 1a. We use the MATLAB implementation of GMRES [17] to solve the saddle point systems (4) for selected ADI shifts q_i that appear during the Newton-ADI iteration. Each q_i is used for

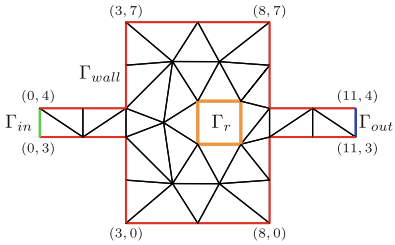


Fig. 1. Initial triangulation of the reactor model with coordinates and boundary conditions [2].

Table 1. Test parameter settings.

Set	Re	Sc
I	1	1
II	1	10
III	10	1
IV	1	100
V	10	10

(a) Different parameter settings.

Variable	Dimension
$n_{\mathbf{z}}$	9 092
$n_{\mathbf{p}}$	1 276
$n_{\mathbf{c}}$	1 187
n	11 555

(b) Different dimensions of FE space.

three ADI steps with four right hand sides every time. Thereby, the number of GMRES steps and the CPU times are measured and arithmetically averaged. The preconditioner $\mathbf{P}$ is evaluated as a MATLAB function handle that solves the linear system (7) using the steps (8). The GMRES tolerance is set to 10^{-10} to ensure the same convergence of the ADI iteration that a direct solve would imply [3]. Although a few complex ADI-shifts q_i appear for each parameter set during the Newton-ADI process, the pictures only show the real parts of q_i.

All computations were executed in MATLAB R2012a on a 64-bit server with 2×Intel® Xeon® X5650 @2.67 GHz, 12 Cores (6 Cores per CPU) and 48 GB main memory available.

4.1 Influence of ADI Shifts and Reynolds and Schmidt Numbers

The influence of the variation of the Reynolds and Schmidt numbers as given in Table 1 is depicted in Fig. 2. To obtain the best approximations for the preconditioning steps (8a)–(8c), a direct solver is used to solve with $F_{\mathbf{z}}, F_{\mathbf{c}}$, and $S_{\mathbf{p}}$. It can be observed that for ADI shifts $-10^5 < \mathrm{Re}\,(q_i) < -10^1$, between 20–25 GMRES steps are needed. As soon as the absolute value of q_i gets smaller then 10 the number of steps increases. This is a natural behavior, because the influence of the mass matrices $M_{\mathbf{z}}$ and $M_{\mathbf{p}}$ vanishes. Nevertheless, GMRES converges within at most 40–80 steps for all parameter configurations. An empirical test to set: $q_i = -10 \quad \forall |q_i| < 10$, during the Newton-ADI process showed similar ADI convergence behavior as for the original shift selection, without the drawback of

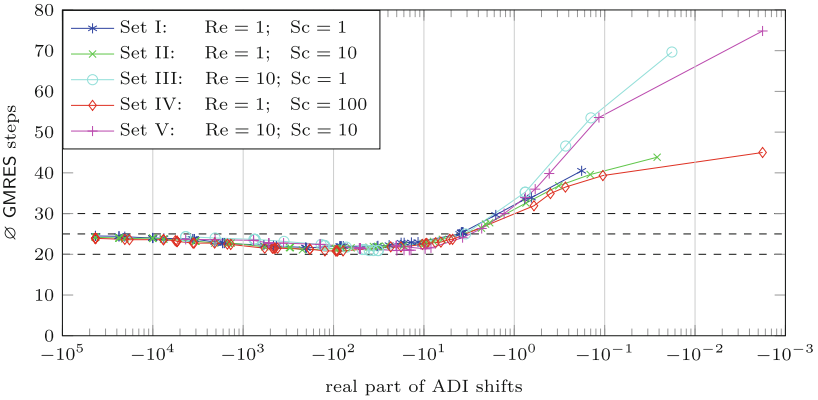


Fig. 2. Average number of GMRES steps for a representative selection of ADI shifts from the Newton-ADI iteration for the configuration sets in Table 1a.

higher GMRES cost for certain shifts. In summary, the derived preconditioner is suitable concerning different Reynolds and Schmidt numbers, as well as different ADI shifts.

4.2 Approximations Using AMG and Chebyshev-Semi-Iteration

As described in Subsect. 3.3, the different preconditioning steps should be solved by an easy to evaluate approximation that is accurate enough to ensure the convergence of GMRES, but avoids the use of sparse factorizations of large-scale sparse matrices. We exchanged the direct solver by its approximation step by step and depict the results in Fig. 3. At first, we use the MATLAB based function `agmg` [11] to solve with $F_{\mathbf{z}}$ and $F_{\mathbf{c}}$ in (8b) and (8a) with an accuracy of 10^{-10}. Depending on the used ADI shift, the function `agmg` needed 1–30 steps. Thus, the times to solve the whole saddle point system with the same number of GMRES steps increased a little bit compared to the direct solver. At second, we approximately solved with $S_{\mathbf{p}}$ in step IIIa using `agmg` as a preconditioner. This was sufficient enough to achieve the GMRES accuracy and, furthermore, decreased the time. Finally, we applied a Chebyshev-Semi-Iteration to approximately solve with $M_{\mathbf{p}}$ in step IIIc. The obtained speedup finally decreased the times below the time used by the direct solver in each step without the loss of any accuracy in GMRES. Due to the above addressed problems with complex ADI shifts in `agmg`, we restrict our comparison in Fig. 3 to a selection of real ADI shifts. The selection has been performed such that the span of all ADI shifts appearing in the entire Newton-ADI process is covered. Where those shifts clustered we only chose one representative per cluster.

At the end of this section it should be noted that the suggested preconditioned GMRES method for the considered class of saddle point problems would show its full strength in comparison to a direct solver when using finer discretizations,

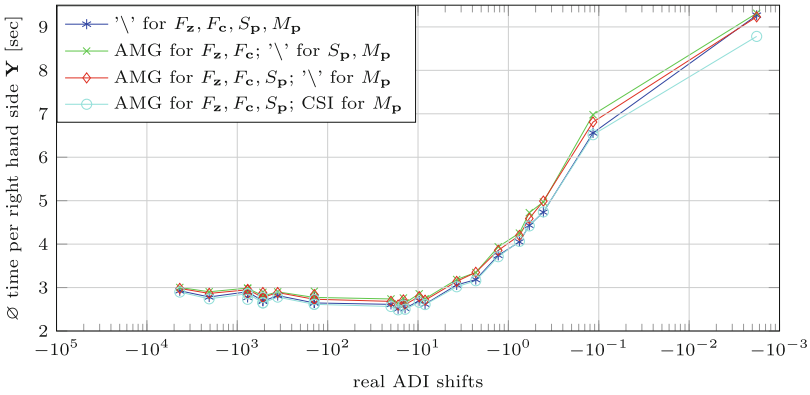


Fig. 3. Average time to solve Eq. (4) with GMRES for a representative selection of real ADI shifts from the Newton-ADI iteration for different approximations of the preconditioning steps (8).

leading to larger dimensions, and in particular when moving to 3D problems. This will be addressed in future work.

5 Conclusions and Outlook

We have recalled the formation of block structured saddle point systems as they arise within the Riccati-based feedback stabilization approach for coupled flow problems that avoids any explicit projection [2]. We were able to extend the results from [3], developed for uncoupled Stokes flow, to the coupled flow described by incompressible Navier-Stokes and a diffusion-convection equation. For that reason, the least-squares commutator approach in [7] has been modified to approximate the shifted Navier-Stokes Schur complement. Exploiting the block structure of the arising preconditioner guarantees a fast evaluation within GMRES.Each of the blocks can either be approximated by an AMG method or a Chebyshev-Semi-Iteration. Several numerical experiments showed that the derived preconditioning method is able to solve the arising saddle point systems efficiently independent of the different parameter settings. Only the use of complex ADI shifts during the Newton-ADI process is not yet optimally covered by this approach and will be investigated in the future.

References

1. Bänsch, E., Benner, P., Saak, J., Weichelt, H.K.: Riccati-based boundary feedback stabilization of incompressible Navier-Stokes flow. Preprint SPP1253-154, DFG-SPP1253 (2013)
2. Bänsch, E., Benner, P., Saak, J., Weichelt, H.K.: Optimal control-based feedback stabilization of multi-field flow problems. In: Leugering, G., Benner, P., Engell, S., Griewank, A., Harbrecht, H., Hinze, M., Rannacher, R., Ulbrich, S. (eds.) Trends in PDE Constrained Optimization. International Series of Numerical Mathematics. Birkhäuser (2014, to appear)

3. Benner, P., Saak, J., Stoll, M., Weichelt, H.K.: Efficient solution of large-scale saddle point systems arising in Riccati-based boundary feedback stabilization of incompressible Stokes flow. SIAM J. Sci. Comput. **35**(5), S150–S170 (2013)
4. Benzi, M., Olshanskii, M.A., Wang, Z.: Modified augmented Lagrangian preconditioners for the incompressible Navier-Stokes equations. Int. J. Numer. Meth. Fluids **66**(4), 486–508 (2011)
5. Cliffe, K.A., Garratt, T.J., Spence, A.: Eigenvalues of block matrices arising from problems in fluid mechanics. SIAM J. Matrix Anal. Appl. **15**(4), 1310–1318 (1994)
6. Elman, H., Howle, V., Shadid, J., Shuttleworth, R., Tuminaro, R.: Block preconditioners based on approximate commutators. SIAM J. Sci. Comput. **27**(5), 1651–1668 (2006)
7. Elman, H., Silvester, D., Wathen, A.: Finite Elements and Fast Iterative Solvers: With Applications in Incompressible Fluid Dynamics. Oxford University Press, Oxford (2005)
8. Heinkenschloss, M., Sorensen, D.C., Sun, K.: Balanced truncation model reduction for a class of descriptor systems with application to the Oseen equations. SIAM J. Sci. Comput. **30**(2), 1038–1063 (2008)
9. Hood, P., Taylor, C.: Navier-Stokes equations using mixed interpolation. In: Oden, J.T., Gallagher, R.H., Taylor, C., Zienkiewicz, O.C. (eds.) Finite Element Methods in Flow Problems, pp. 121–132. University of Alabama in Huntsville Press, Huntsville (1974)
10. Napov, A., Notay, Y.: An algebraic multigrid method with guaranteed convergence rate. SIAM J. Sci. Comput. **34**, A1079–A1109 (2012)
11. Notay, Y.: An aggregation-based algebraic multigrid method. ETNA, Electron. Trans. Numer. Anal. **37**, 123–146 (2010)
12. Notay, Y.: Aggregation-based algebraic multigrid for convection-diffusion equations. SIAM J. Sci. Comput. **34**, A2288–A2316 (2012)
13. Raymond, J.P.: Feedback boundary stabilization of the two-dimensional Navier-Stokes equations. SIAM J. Control Optim. **45**(3), 790–828 (2006)
14. Ruge, J., Stüben, K.: Algebraic multigrid (amg). In: McCormich, E. (ed.) Multigrid Methods. Frontiers in Applied Mathematics, vol. 5, pp. 73–130. SIAM, Philadelphia (1987)
15. Saad, Y.: A flexible inner-outer preconditioned gmres algorithm. SIAM J. Sci. Comput. **14**(2), 461–469 (1993)
16. Saad, Y.: Iterative Methods for Sparse Linear Systems. SIAM, Philadelphia (2003)
17. Saad, Y., Schultz, M.H.: GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. **7**(3), 856–869 (1986)
18. Stoll, M., Wathen, A.: All-at-once solution of time-dependent Stokes control. J. Comp. Phys. **232**, 498–515 (2013)

Stochastic Control of Econometric Models for Slovenia

Dimitri Blueschke, Viktoria Blueschke-Nikolaeva, and Reinhard Neck(✉)

Department of Economics, Alpen-Adria-Universität Klagenfurt,
Universitätsstraße 65-67, 9020 Klagenfurt, Austria
reinhard.neck@uni-klu.ac.at

Abstract. This paper considers the optimal control of two small stochastic models of the Slovenian economy applying the OPTCON algorithm. OPTCON determines approximate numerical solutions to optimum control problems for nonlinear stochastic systems and is particularly applicable to econometric models. We compare the results of applying the OPTCON2 version of the algorithm to the nonlinear model SLOVNL and the linear model SLOVL. The results for both models are similar, with open-loop feedback controls giving better results on average but with more 'outliers' than open-loop controls.

Keywords: Optimal control · Stochastic control · Algorithms · Policy applications · Nonlinear models

1 Introduction

Optimum control theory has been applied in many areas of science, from engineering to economics. An algorithm that provides (approximate) solutions to optimum control problems for nonlinear dynamic systems with different kinds of stochastics is OPTCON, which was first introduced in [3]. An extension has been developed in [2], which includes passive learning or open-loop feedback control policies.

OPTCON was implemented in MATLAB and can deliver numerical solutions to problems with real economic data. Two such applications are described and analyzed in this paper. We develop two macroeconomic models of the Slovenian economy, a nonlinear model called SLOVNL and a (comparable) linear model called SLOVL. The algorithm with both open-loop and open-loop feedback strategies is applied to these models and the influence of the control scheme and the nonlinearity of the model on the optimum solution is investigated in some optimization experiments.

2 The Problem

The OPTCON algorithm is designed to provide approximate solutions to optimum control problems with a quadratic objective function (a loss function to

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 21–30, 2014.
DOI: 10.1007/978-3-662-45504-3_3

be minimized) and a nonlinear multivariate discrete-time dynamic system under additive and parameter uncertainties. The intertemporal objective function is formulated in quadratic tracking form, which is quite often used in applications of optimum control theory to econometric models. It can be written as

$$J = E\left[\sum_{t=1}^{T} L_t(x_t, u_t)\right], \tag{1}$$

with

$$L_t(x_t, u_t) = \frac{1}{2}\begin{pmatrix} x_t - \tilde{x}_t \\ u_t - \tilde{u}_t \end{pmatrix}' W_t \begin{pmatrix} x_t - \tilde{x}_t \\ u_t - \tilde{u}_t \end{pmatrix}. \tag{2}$$

x_t is an n-dimensional vector of state variables that describes the state of the economic system at any point in time t. u_t is an m-dimensional vector of control variables, $\tilde{x}_t \in R^n$ and $\tilde{u}_t \in R^m$ are given 'ideal' (desired, target) levels of the state and control variables respectively. T denotes the terminal time period of the finite planning horizon. W_t is an $((n+m) \times (n+m))$ matrix, specifying the relative weights of the state and control variables in the objective function. W_t (or W) is symmetric.

The dynamic system of nonlinear difference equations has the form

$$x_t = f(x_{t-1}, x_t, u_t, \theta, z_t) + \varepsilon_t, \quad t = 1, ..., T, \tag{3}$$

where θ is a p-dimensional vector of parameters the values of which are assumed to be constant but unknown to the decision maker (parameter uncertainty), z_t denotes an l-dimensional vector of non-controlled exogenous variables, and ε_t is an n-dimensional vector of additive disturbances (system error). θ and ε_t are assumed to be independent random vectors with expectations $\hat{\theta}$ and O_n respectively and covariance matrices $\Sigma^{\theta\theta}$ and $\Sigma^{\varepsilon\varepsilon}$ respectively. f is a vector-valued function fulfilling some differentiability assumptions, $f^i(.....)$, is the i-th component of $f(.....)$, $i = 1, ..., n$.

3 The Optimum Control Algorithm

The OPTCON1 algorithm [3] determines policies belonging to the class of open-loop controls. It either ignores the stochastics of the system altogether or assumes the stochastics to be given once and for all at the beginning of the planning horizon. The nonlinearity problem is tackled iteratively, starting with a tentative path of state and control variables. The tentative path of the control variables is given for the first iteration. In order to find the corresponding tentative path for the state variables, the nonlinear system is solved numerically. After the tentative path is found, the iterative approximation of the optimal solution starts. The solution is sought from one time path to another until the algorithm converges or the maximal number of iterations is reached. During this search the system

is linearized around the previous iteration's result as a tentative path and the problem is solved for the resulting time-varying linearized system. The criterion for convergence demands that the difference between the values of current and previous iterations be smaller than a pre-specified number. The approximately optimal solution of the problem for the linearized system is then used as the tentative path for the next iteration, starting off the procedure all over again.

The more recent version OPTCON2 [2] incorporates both open-loop and open-loop feedback (passive-learning) controls. The idea of passive learning corresponds to actual practice in applied econometrics: at the end of each time period, the model builder (the control agent) observes what has happened, that is, the current values of state variables, and uses this information to re-estimate the model and hence improve his/her knowledge of the system.

The passive-learning strategy implies observing current information and using it in order to adjust the optimization procedure. For the purpose of comparing open-loop and open-loop feedback results, it is not possible to observe current and true values, so one has to resort to Monte Carlo simulations. Large numbers of random time paths for the additive and multiplicative errors are generated, representing what new information could look like in reality. In this way 'quasi-real' observations are created and both types of controls, open-loop and passive-learning (open-loop feedback), are compared.

4 The SLOVNL Model

We estimated two simple macroeconometric models for Slovenia, one nonlinear (SLOVNL) and one linear (SLOVL). The SLOVNL model (**SLOV**enian model, **N**on-**L**inear version) is a small nonlinear econometric model of the Slovenian economy consisting of 8 equations, 4 behavioral equations and 4 identities. SLOVNL includes 8 state variables, 3 control variables, 4 exogenous non-controlled variables and 16 unknown (estimated) parameters. We used quarterly data for the time periods 1995:1 to 2006:4; this data base with 48 observations admits a full-information maximum likelihood (FIML) estimation of the expected values and the covariance matrices for the parameters and the system errors. The starting period for the optimization is 2004:1; the terminal period is 2006:4 (12 periods).

Model variables used in SLOVNL:

Endogenous (state) variables:

$x[1]$	CR	real private consumption
$x[2]$	$INVR$	real investment
$x[3]$	$IMPR$	real imports of goods and services
$x[4]$	$STIRLN$	short term interest rate
$x[5]$	$GDPR$	real gross domestic product
$x[6]$	VR	real total aggregate demand
$x[7]$	PV	general price level
$x[8]$	$Pi4$	rate of inflation

Control variables:

u[1]	$TaxRate$	net tax rate
u[2]	GR	real public consumption
u[3]	$M3N$	money stock, nominal

Exogenous non-controlled variables:

z[1]	EXR	real exports of goods and services
z[2]	$IMPDEF$	import price level
z[3]	$GDPDEF$	domestic price level
z[4]	$SITEUR$	nominal exchange rate SIT/EUR

Model equations:
Standard deviations are given in brackets.

$$\begin{array}{l} CR_t = \underset{(189.7449)}{240.9398} + \underset{(0.1115)}{0.740333}\, CR_{t-1} + \underset{(0.0330)}{0.111727}\, GDPR_t \left(1 - \frac{TaxRate_t}{100}\right) \\ \quad - \underset{(2.5848)}{1.007353}\, (STIRLN_t - Pi4_t) - \underset{(2.4966)}{4.773533}\, Pi4_t \end{array}$$

$$\begin{array}{l} INVR_t = \underset{(176.8549)}{75.41731} + \underset{(0.1423)}{0.932211}\, INVR_{t-1} + \underset{(0.0924)}{0.264523}\, (VR_t - VR_{t-1}) \\ \quad - \underset{(6.9044)}{0.455511}\, (STIRLN_t - Pi4_t) - \underset{(3.1277)}{2.981241}\, Pi4_t \end{array}$$

$$IMPR_t = IMPR_{t-1} + \underset{(0.0724)}{0.826449}\, (VR_t - VR_{t-1}) - \underset{(18.9336)}{38.14954}\, SITEUR_t$$

$$\begin{array}{l} STIRLN_t = \underset{(0.1375)}{0.811606}\, STIRLN_{t-1} - \underset{(0.0008)}{0.000877}\, \frac{(M3N)_t}{PV_t} \cdot 100 \\ \quad + \underset{(0.0026)}{0.002746}\, GDPR_t \end{array}$$

$$GDPR_t = CR_t + INVR_t + GR_t + EXR_t - IMPR_t$$

$$VR_t = GDPR_t + IMPR_t$$

$$PV_t = \frac{GDPR_t}{VR_t} \cdot GDPDEF_t + \frac{IMPR_t}{VR_t} \cdot IMPDEF_t$$

$$Pi4_t = \frac{PV_t - PV_{t-4}}{PV_{t-4}} \cdot 100$$

The objective function penalizes deviations of objective variables from their 'ideal' (desired, target) values. The 'ideal' values of the state and control variables ($\tilde{x}_t$ and $\tilde{u}_t$ respectively) are chosen as shown in Table 1. The 'ideal' values for most variables are defined in terms of growth rates (denoted by % in Table 1) starting from the last given observation (2003:4). For $Pi4$ and $TaxRate$, constant 'ideal' values are used; for $STIRLN$, a linear decrease of 0.25 per quarter is assumed to be the goal.

Table 1. 'Ideal' values of objective variables, SLOVNL

CR	$INVR$	$IMPR$	$STIRLN$	$GDPR$	VR	PV	$Pi4$	$TaxRate$	GR	$M3N$
1%	1%	2%	-0.25	1%	1.5%	0.75%	3	25.2	1%	1.75%

The weights for the variables, i.e. the constant matrix W in the objective function, are first chosen as shown in Table 2a ('raw' weights) to reflect the relative importance of the respective variable in the (hypothetical) policy maker's

Table 2. Weights of objective variables, SLOVNL

2a: 'raw' weights

variable	weight
CR	1
$INVR$	1
$IMPR$	1
$STIRLN$	1
$GDPR$	2
VR	1
PV	1
$Pi4$	0
$TaxRate$	2
GR	2
$M3N$	2

2b: 'correct' weights

variable	weight
CR	3.457677
$INVR$	12.16323
$IMPR$	1.869532
$STIRLN$	216403.9
$GDPR$	2
VR	0.333598
PV	423.9907
$Pi4$	0
$TaxRate$	37770.76
GR	63.77052
$M3N$	0.090549

objective function. These 'raw' weights have to be scaled or normalized according to the levels of the respective variables to make the weights comparable. The normalized ('correct') weights are shown in Table 2b.

5 The SLOVL Model

To analyse the impact of the nonlinearity of the system we developed a linear pendant to the SLOVNL model. This 'sister model' is called SLOVL (**SLOV**enian model, **L**inear version) and consists of 6 equations, 4 of them behavioral and 2 identities. The model includes 6 state variables, 3 exogenous non-controlled variables, 3 control variables, and 15 unknown (estimated) parameters. We used the same data base as for SLOVNL and a specification as close as possible to that of SLOVNL in order to make comparisons between the results of the algorithm for a linear and a nonlinear model. Again, we used full-information maximum likelihood (FIML) to estimate the expected values and the covariance matrices for the parameters and the system errors. The starting and the terminal period for the optimization are again 2004:1 and 2006:4.

Model variables used in SLOVL:

Endogenous (state) variables:

$x[1]$	CR	real private consumption
$x[2]$	$INVR$	real investment
$x[3]$	$IMPR$	real imports of goods and services
$x[4]$	$STIRLN$	short term interest rate
$x[5]$	$GDPR$	real gross domestic product
$x[6]$	VR	real total aggregate demand

Control variables:

u[1]	$Taxes$	tax revenue
u[2]	GR	real public consumption
u[3]	$M3R$	money stock, real

Exogenous non-controlled variables:

z[1]	EXR	real exports of goods and services
z[2]	$SITEUR$	nominal exchange rate SIT/EUR
z[3]	$Pi4$	rate of inflation

Model equations:
Standard deviations are given in brackets.

$$\begin{array}{l} CR_t = 231.582776 + 0.744522\ CR_{t-1} + 0.111736\ (GDPR_t - Taxes_t) \\ \qquad\quad (191.99) \qquad\quad (0.11) \qquad\quad (0.03) \\ \qquad - 0.855137\ (STIRLN - Pi4) - 4.657411\ Pi4 \\ \qquad\quad (2.63) \qquad\quad (2.50) \end{array}$$

$$\begin{array}{l} INVR_t = 69.965212 + 0.936305\ INVR_{t-1} + 0.265119\ (VR_t - VR_{t-1}) \\ \qquad\quad (176.51) \qquad\quad (0.14) \qquad\quad (0.09) \\ \qquad - 0.292918\ (STIRLN_t - Pi4_t) - 2.869522\ Pi4_t \\ \qquad\quad (6.90) \qquad\quad (3.11) \end{array}$$

$$\begin{array}{l} IMPR_t = IMPR_{t-1} + 0.826648\ (VR_t - VR_{t-1}) - 38.158117\ SITEUR_t \\ \qquad\quad (0.07) \qquad\quad (18.86) \end{array}$$

$$\begin{array}{l} STIRLN_t = 0.811458\ STIRLN_{t-1} - 0.000877\ (M3R)_t \\ \qquad\quad (0.14) \qquad\quad (0.0008) \\ \qquad + 0.002748\ GDPR_t \\ \qquad\quad (0.0026) \end{array}$$

$$GDPR_t = CR_t + INVR_t + GR_t + EXR_t - IMPR_t$$

$$VR_t = GDPR_t + IMPR_t$$

The objective function is analogous as for SLOVNL, where the 'ideal' values of the state and control variables ($\tilde{x}_t$ and $\tilde{u}_t$ respectively) are chosen as shown in Table 3. For the weights for the variables, Table 4a shows the 'raw' weights and Table 4b gives the normalized weights.

Table 3. 'Ideal' values of objective variables, SLOVL

CR	$INVR$	$IMPR$	$STIRLN$	$GDPR$	VR	$Taxes$	GR	$M3R$
1%	1%	2%	-0.25	1%	1.5%	25.2	1%	1.75%

Table 4. Weights of objective variables, SLOVL

4a: 'raw' weights

variable	weight
CR	1
$INVR$	1
$IMPR$	1
$STIRLN$	1
$GDPR$	2
VR	1
$Taxes$	2
GR	2
$M3R$	2

4b: 'correct' weights

variable	weight
CR	3.457677
$INVR$	12.16323
$IMPR$	1.869532
$STIRLN$	216403.9
$GDPR$	2
VR	0.333598
$Taxes$	27.52906
GR	63.77052
$M3R$	0.292662

6 Optimization Experiments

The OPTCON2 algorithm is applied to the two econometric models SLOVNL and SLOVL. Two different experiments are run for both models: in experiment 1, two open-loop solutions are compared, a deterministic one where the variances and covariances of the parameters are ignored, and a stochastic one where the estimated parameter covariance matrix is taken into account. In experiment 2, the properties of the open-loop and the open-loop feedback solutions are compared. Furthermore, by comparing the results for the SLOVNL and the SLOVL models we want to analyze the impact of nonlinearity on the properties of the optimal solution.

6.1 Experiment 1: Open-Loop Optimal Policies

For experiment 1, two different open-loop solutions are calculated: a deterministic and a stochastic one. The deterministic solution assumes that all parameters of the model are known with certainty and are equal to the estimated values. In the stochastic case, the covariance matrix of the parameters as estimated by FIML is used but no updating of information occurs during the optimization process.

The results (for details, see [1,2]) show that both the deterministic and the stochastic solutions follow the 'ideal' values fairly well but fiscal policies are less expansionary and hence real GDP is mostly below its 'ideal' values. The values of the objective function show a considerable improvement in system performance obtained by optimization and only moderate costs of uncertainty.

An interesting result is that the deterministic and the stochastic open-loop solutions are very similar. Furthermore, one can see that the SLOVL model is a good 'linear approximation' of the nonlinear SLOVNL model because the results for both models are nearly identical. This fact can be used for isolating the impact of nonlinearity on finding the optimum control solution, especially for the case of open-loop feedback policies.

6.2 Experiment 2: Open-Loop Feedback Optimal Policies

The aim of experiment 2 consists in comparing open-loop (OL) and open-loop feedback (OLF) optimal stochastic controls. Figures 1 and 2 show the results of a representative Monte Carlo simulation, displaying the value of the objective function arising from applying OPTCON2 to the SLOVNL and the SLOVL models respectively, under 1000 independent random Monte Carlo runs. The graphs plot the values of the objective function for OL policies (x-axis) and OLF policies (y-axis) against each other. In the 'zoom in' panels of the figures, we cut the axes so as to show the mass of the points and omitting 'outliers', i.e. results where the value of the objective function becomes extremely large.

One can see that in most cases the values of the objective function for the open-loop feedback solution are smaller than the values of the open-loop solution, indicated by a greater mass of dots below the 45 degree line. This means that

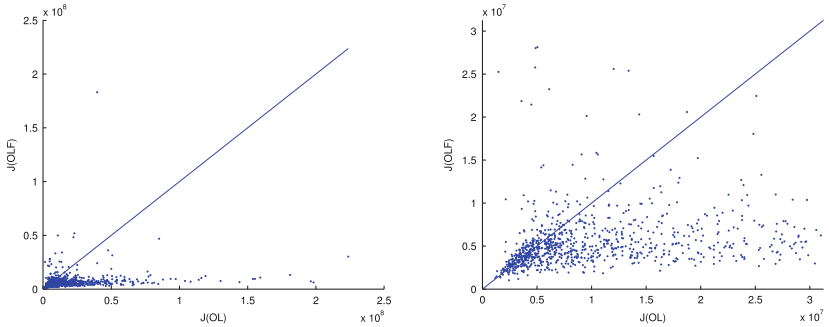


Fig. 1. OL and OLF control, value of objective function; SLOVNL; 1000 Monte Carlo runs; left: 'normal', right: 'zoom in'

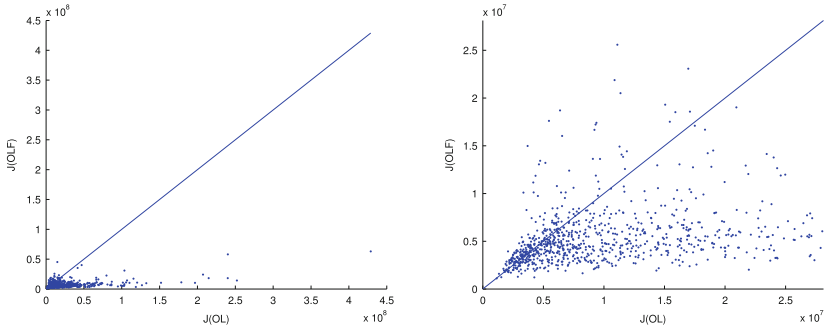


Fig. 2. OL and OLF control, value of objective function; SLOVL; 1000 Monte Carlo runs; left: 'normal', right: 'zoom in'

open-loop feedback controls give better results (lower values of the cost function) in the majority of the cases investigated. For the SLOVNL model, the OLF policy gives better results than the OL policy in 66.4 % of the cases, for the SLOVL model in 65.4 % of the cases considered here.

However, one can also see from these figures (especially in the left-hand panels with a 'normal' view) that there are many cases where either control scheme results in very high losses, indicated by dots which are significantly distant from the main mass of the dots. These cases are called 'outliers' and can be seen even more clearly in Fig. 3. This figure shows the same results of the 1000 independent Monte Carlo runs for each model (SLOVNL and SLOVL) separately, but for each Monte Carlo run. The OLF and OL objective function values are plotted in Fig. 3 together on the y-axis in each Monte Carlo run, the number of which is shown on the x-axis. Diamonds represent open-loop feedback results and squares represent open-loop results.

The results mean that (passive) learning does not necessarily improve the quality of the final results; it may even worsen them. One reason for this is the presence of the two types of stochastic disturbances: additive (random system

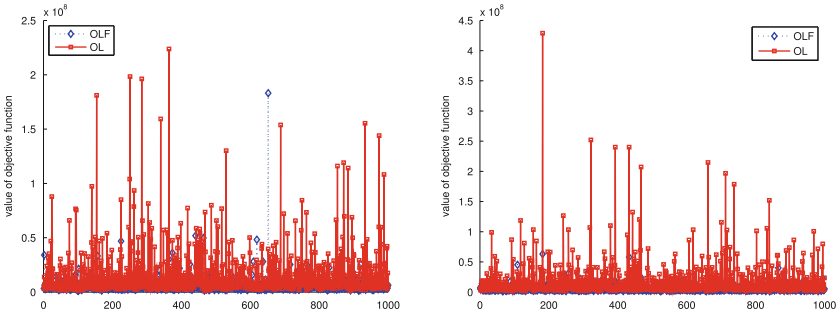


Fig. 3. Open-loop vs. open-loop feedback control, value of objective function (1000 Monte Carlo runs) (left: SLOVNL, right: SLOVL)

error) and multiplicative error ('structural' error in the parameters). The decision maker cannot distinguish between realizations of errors in the parameters and in the equations as he just observes the resulting state vector. Based on this information, he learns about the values of the parameter vector, but he may be driven away from the 'true' parameter vector due to the presence of the random system error.

6.3 On the Impact of Nonlinearity

In the previous section we saw that there is a severe problem of what we called 'outliers' - cases with very high losses or values of the objective function. In a similar framework of optimum stochastic control, [4] investigated numerical reasons for outliers. It was not possible to confirm that the sources of the problem found by these authors were decisive for the outlier problem in our framework. We suspect that there are other reasons for the outliers. One possible reason is the stochastics of the dynamic system itself. In our SLOVNL and SLOVL models all the parameters (including all the intercepts) are considered to be stochastic, which may make this reason more likely to work.

The second possible reason is based on the nonlinear nature of the models for which the OPTCON algorithm was created. The SLOVL model was created mainly in order to test this possibility. The graphical results in the previous section show that the outliers occur in the linear as well as in the nonlinear model version. Moreover, in some of the experiments with 1000 Monte Carlo runs for the SLOVNL model, it turned out that the algorithm did not converge in some runs. In these cases, the algorithm starts to diverge and results in some non-reasonable or even complex numbers for some variables. In the 1000 Monte Carlo runs experiment considered above this happened six times. On the contrary, under the SLOVL model, not one single case of non-convergence out of the 1000 Monte Carlo runs occurred. Thus we arrive at the conclusion that nonlinearity is not the reason for the 'outliers', but it can worsen the problem.

7 Conclusion

A comparison of open-loop control (without learning) and open-loop feedback control (with passive learning) shows that open-loop feedback control outperforms open-loop control in the majority of the cases investigated for the two small econometric models of Slovenia. But it suffers from a problem of 'outliers' which is present for both policy schemes. When comparing the results for the nonlinear SLOVNL model and the linear SLOVL model, we found that the nonlinearity of the system is not responsible for the 'outliers' but may worsen their influence in some cases.

References

1. Blueschke, D., Blueschke-Nikolaeva, V., Neck, R.: Stochastic control of linear and nonlinear econometric models: some computational aspects. Comput. Econ. **42**, 107–118 (2013)
2. Blueschke-Nikolaeva, V., Blueschke, D., Neck, R.: Optimal control of nonlinear dynamic econometric models: an algorithm and an application. Comput. Stat. Data Anal. **56**, 3230–3240 (2012)
3. Matulka, J., Neck, R.: OPTCON: an algorithm for the optimal control of nonlinear stochastic models. Ann. Oper. Res. **37**, 375–401 (1992)
4. Tucci, M.P., Kendrick, D.A., Amman, H.M.: The parameter set in an adaptive control monte carlo experiment: some considerations. J. Econ. Dyn. Control **34**, 1531–1549 (2010)

The Optimal Control of Cellular Communication Enterprise Development in Competitive Activity

Irina Bolodurina(✉) and Tatyana Ogurtsova

Orenburg State University, Pobedi 13, Orenburg 460018, Russia
prmat@mail.osu.ru
http://www.osu.ru

Abstract. The work is devoted to the construction and justification of the mathematical model of the competitive behaviour of cellular communication in the form of a system of nonlinear differential equations with delay time describing the dynamics of changes in the subscriber base of cellular operators competing for shared resources.

Keywords: Mathematical model · Differential equations with lag · Optimal control · Identification of parameters · Principal of pontryagins maximum

Competition is essential and most important in many actual processes. Mathematical modeling is a basic way to control the agents of competition. It allows to consider the different factors influencing their interaction dynamics. Actual agent competitive behavior modeling widely uses nonlinear differential equations with retarded argument which provides more complex structure dynamic models. That is why the development of bundled software based on efficient numerical methods aimed to solve nonlinear dynamics and control problems is an essential scientific task.

There are many works on nonlinear object control. But the task to determine optimal control has not been studied for enterprise development dynamic models with a phase variable lag. The urgent objective here is to find efficient algorithms of optimal control solutions in such systems as well as to develop the adequate software. Thus the management of enterprises will be provided with Decision Support Systems aimed to develop the communication service tariff policy.

There are three cellular operators in Russia. They are: MTS, Beeline and Megafon. There are also many smaller cellular enterprises. The 2004–2010 data of user base and tariff policy have been used for our model parameter identification. The research has been carried out based on the interaction of two economic agents. We divide all the operators into two unequal groups. The first economic agent (EA1) is one of the leading cellular enterprises. The other economic agent (EA2) includes all the rival enterprises summing the number of their users in the actual market.

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 31–35, 2014.
DOI: 10.1007/978-3-662-45504-3_4

Assume with no competition the number of customers for each economic agent grows exponentially, with increment rate ε_i. Considering the saturation effect in the cellular service market as well as competition we describe the development dynamics for the two agents user base through the system of differential equations

$$\dot{x_i}(t) = x_i(t) \left[\varepsilon_i - \sum_{k=1}^{2} \gamma_{ik} x_k(t) \right]. \tag{1}$$

The time lag is to be considered in interaction of economic agents as the time difference between the moment of managerial decision and the actual changes of market situation. The user base general dynamics can be expressed through the system of nonlinear differential equations with retarded argument

$$\dot{x_i}(t) = x_i(t) \left[\varepsilon_i - \sum_{k=1}^{2} \gamma_{ik} x_k(t-\tau) \right]. \tag{2}$$

To model the cellular enterprise behavior control we introduce an intensity brackets $u_i(t), i = 1, 2$ thus describing the average minute cost at moment t and satisfying the restriction

$$\alpha_i \le u_i(t) \le \beta_i, \ i = 1, 2, \ t \in [0, T]. \tag{3}$$

Thus the controllable model of two economic agent interaction in the conditions of competition for cellular service users can be described through Lotka-Volterra logistic model with retarded argument

$$\begin{aligned} \dot{x}_1(t) &= x_1(t) \left[\varepsilon_1 - \gamma_{11} x_1(t-\tau) - \gamma_{12} x_2(t-\tau) \right] - p_{11} u_1(t) - p_{12} u_2(t), \\ \dot{x}_2(t) &= x_2(t) \left[\varepsilon_2 - \gamma_{21} x_1(t-\tau) - \gamma_{22} x_2(t-\tau) \right] - p_{21} u_1(t) - p_{22} u_2(t). \end{aligned} \tag{4}$$

There is a lower boundary of user base volume providing normal operation, as well as an upper boundary determined by the network performance specifications. The above are presented through phase variable restrictions

$$\eta_i \le x_i(t) \le \mu_i, \ i = 1, 2. \tag{5}$$

The number of cellular Operator I users at the initial interval $[-\tau, 0]$ is determined by the functions

$$x_i(t) = \varphi_i(t), \ t \in [-\tau, 0]. \tag{6}$$

Parameter identification for the system of differential equations is an important procedure in modeling and solution of optimal control problem. Based on the method of model setting to the experimentally obtained data, the approach to the simultaneous identification of system parameters and the value of the lag is the subject of our research. The parameters are set to minimize the functionals characterizing the quality of model settings.

This approach implies multiple monitoring of the dynamic process, considerably exceeding the dimensionality of the system.

In course of economic agent interaction a company may face different managerial goals expressed in corresponding quality criteria (7) (9). According to the first economic agents (EA1) development priorities we find the optimal control solution for each objective functional:

1. user base growth

$$J_1(u_1) = \int_0^T b_1 x_1(t) dt \to \max; \tag{7}$$

2. reaching the aimed user base

$$J_2(u_1) = b_2 \left(x_1(T) - M\right)^2 \to \min, \tag{8}$$

where M – planning value of the subscriber base EA1;

3. EA1 income growth

$$J_3(u_1) = \int_0^T b_3 x_1(t) u_1(t) dt \to \max. \tag{9}$$

The problems set above can be classified as optimal control fixed lag problems. To solve them, in this research we have used Pontryagin's maximum principle for fixed-lag systems. We suggested that the cost value of one EA2 minute is fixed and can be evaluated by the previous tariff policy dynamics.

The optimal control problem is characterized by: Nonlinearity of the differential equation system describing the user base dynamics for the two economical agents; constant lag in the controlled systems state vector. Considering the above features we obtained the optimality conditions. We based the solution numerical algorithm on that. The control restrictions in the problem have been considered due to the gradient projection method, given an arbitrary choice of control initial approximation.

The use of one numerical method is not always enough to find the solution for the optimal control problem with the required precision. That is why different algorithms can be used in the process of solution. One of the approximate methods used to solve nonlinear object optimal control problems was suggested by L.I. Shatrovsky. Its based on linearization of a set non-linear system and further iterative procedure solving a linear problem at each step so approximating the initial task. As a result we get a control close enough to the optimal. To improve the calculation reliability solving the optimal control problems we suggest a combined method. It implies the choice of Shatrovsky method allowable control as the initial approximation in the gradient projection method. This approach to the choice of control initial approximation allows to avoid the functional in the local extremum.

Our methods and algorithms have served as the basis for bundled software aimed at the search of numerical solution for an optimal control problem, two economical agents competing for communication users. This software finds the index values for the parameters and lag by the user base statistics data and the previous tariff policy. It also proves helpful for communication enterprise managerial strategies according to development priorities.

Computing experiments have been carried out to examine the efficiency of the suggested models and algorithms for the major communication operators in the Russian market.

The conclusion has been made that the introduction of control and lag in the cellular communication enterprise model improves the values of the objective functional which substantiates the choice of the controlled system (9)

$$\dot{x}_1(t) = x_1(t)\left[0.294 - 0.0048x_1(t) - 0.00089x_2(t)\right] - 1.1681u_1(t) + 2.672u_2(t),$$
$$\dot{x}_2(t) = x_2(t)\left[0.193 + 0.0043x_1(t) - 0.0034x_2(t)\right] + 0.105u_1(t) + 0.424u_2(t).$$

The established relation is efficient approximating the experimental data, the difference between calculation data and experimental data being 6,5%. User base and tariff policy updates are followed up by parameter identification which enhances modelling precision.

Based on the gradient projection method and combined method we determine the optimal decision for each objective functional as well as find the corresponding values.

Comparative analysis shows that the control achieved with the use of the suggested combined method is more efficient to increase the EA1 user base.

The EA1 user base values have been calculated for each objective functional at the end of 30 month forecasting interval. A significant growth of EA1 user base by this moment proves the efficiency of suggested combined method for solution of cellular communication enterprise interaction.

	Values $x_1(T)$, obtained with the use of:		Values objective functional, obtained ith the use of:	
	Gradient projection method, million users	Combined method, million users	Gradient projection method, million users	Combined method, million users
User base growth ($J_1(u_1)$)	30.768	31.146	127.02	142.35
Reaching the aimed user base ($J_2(u_1)$)	31.281	31.331	0.169	0.158
Income growth ($J_3(u_1)$)	30.989	31.002	8.78	9.31

Our bundled software has been used by Orenburg branch of ROSTELECOM controlling their tariff policy.

The developed models and algorithms can be modified for any economical agents practical problems in their competitive activity, such as web-site user base or radio listener base control or TV channel rating etc.

References

1. Andreeva, E.A., Tciruleva, V.A.: Calculus Of Variations and Optimization Methods, p. 575. Tver state University, Orenburg, Tver (2004)
2. Bolodurina, I.P.: Differential equations with delay argument and applications. Orenburg. st. un-ty, 101 p. (2006)
3. Bolodurina, I.P., Ogurtsova, T.A.: Managing the price for services of the enterprises of the telecommunications industry. Prob. Manage. **N3**, 30–35 (2011)
4. Koblov, A.I., Shiryaev, V.I.: Optimal control of the behaviour of firms on the market of cellular communication. Theory Syst. **N5**, 157–165 (2008)
5. Prasolov, A.V.: Dynamic models with delay and applications in Economics and engineering. SPb.: Fallow deer, 192 p. (2010)

Simulation of Acoustic Wave Propagation in Anisotropic Media Using Dynamic Programming Technique

Nikolai Botkin(✉) and Varvara Turova

Center for Mathematics, Technical University of Munich, Boltzmannstr. 3,
85748 Garching b. Munich, Germany
{botkin,turova}@ma.tum.de

Abstract. It is known that the Hamiltonian of the eikonal equation for an anisotropic medium may be nonconvex, which excludes the application of Fermat's minimum-time principle related to minimum-time control problems. The idea proposed in this paper consists in finding a conflict control problem (differential game) whose Hamiltonian coincides with the Hamiltonian of the eikonal equation. It turns out that this is always possible due to Krasovskii's unification technique. Having such a differential game allows us to apply dynamic programming methods to computing the value function of the game, and therefore to describe the propagation of wave fronts. This method is very appropriate for the simulation of wave patterns in surface acoustic wave biosensors. Numerical computations given in this paper prove the feasibility of the approach proposed.

Keywords: WKB-approximation · Hamilton-Jacobi equations · Viscosity solutions · Differential game · Unification

1 Introduction

The paper concerns the development of methods for modeling the propagation of acoustic waves in anisotropic media. This investigation is very important for many applications such as acoustic sensors whose operating principle is based on the excitation and detection of acoustic waves of very high frequency in piezoelectric crystals.

For anisotropic media, the WKB (Wentzel-Kramers-Brillouin) approximation yields eikonal equations whose Hamiltonians are neither convex nor concave in the impulse variable. Therefore, the well-known Fermat principle of wave propagation fails in this case. Moreover, the propagation occurs in such a way as if an antagonistic opponent aims to slow down the movement of the wave fronts. Thus, we arrive at the idea to use methods of differential games in the analysis of wave propagation. If the negative of the Hamiltonian of a

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 36–51, 2014.
DOI: 10.1007/978-3-662-45504-3_5

differential game approximates the Hamiltonian of the eikonal equation, then the value function of the game approximates the phase function satisfying the eikonal equation. The authors have developed effective and precise algorithms for solving Hamilton-Jacobi equations arising from differential games, which yields an effective tool for the numerical investigation of eikonal equations.

If a differential game is chosen appropriately, the level sets of its value function represent the wave fronts, and optimal trajectories are associated with the propagation of rays. Thus, it makes possible to describe very complicated behavior of rays using game-theoretic classification of the so-called singular surfaces that can attract, repulse, and break the trajectories. For example, the caustic-like behavior of rays can be interpreted as the attraction of neighboring optimal trajectories to a singular surface. The monograph [1] extends the classical method of characteristics by introducing the so-called generalized characteristics that are related to the above mentioned trajectories and singular surfaces.

The main objective of this paper is the numerical simulation of propagation of bulk and surface acoustic waves in anisotropic monocrystals and multi-layered structures used in surface acoustic wave sensors. It is demonstrated that the propagation fronts can be found very precisely even in the case of very complicated geometry of wave emitters. Numerical results are presented for the case of bulk and surface waves characterized by non-convex slowness surfaces.

This investigation is inspired by the cooperation with professor A. A. Melikyan (deceased) from the Institute for Problems in Mechanics, Moscow, Russia.

2 Wave Velocity in Piezoelectric Crystals and Eikonal Equation

Assume that the indexes i, j, k, l run from 1 to 3 and use the summation convention over the repeated indexes. Let u_1, u_2, and u_3 be the displacements in x_1, x_2, and x_3 directions, respectively; φ is the electric potential such that the electric field E_i is given by the relation $E_i = \partial\varphi/\partial x_i$. Electro-elasticity equations for a piezoelectric anisotropic crystal read:

$$\rho u_{itt} - C_{ijkl}\frac{\partial^2 u_l}{\partial x_j \partial x_k} + e_{kij}\frac{\partial^2 \phi}{\partial x_k \partial x_j} = 0, \tag{1}$$

$$\epsilon_{ij}\frac{\partial^2 \phi}{\partial x_i \partial x_j} + e_{ikl}\frac{\partial^2 u_l}{\partial x_i \partial x_k} = 0\,. \tag{2}$$

where ρ, ϵ_{ij}, e_{ikl}, and C_{ijkl} denote the density, the material dielectric tensor, the stress piezoelectric tensor, and the elastic stiffness tensor, respectively.

The WKB (high frequency) approximation (see e.g. [2]) uses the ansatz

$$u_j = u_j^0(t,x)\cdot\varepsilon e^{\imath S(t,x)/\varepsilon}, \quad \phi = \phi^0(t,x)\cdot\varepsilon e^{\imath S(t,x)/\varepsilon}, \tag{3}$$

where $\varepsilon = \omega^{-1}$ is a small parameter (ω is the frequency), $S(t,x)$ is the phase function, and $u_j^0(t,x)$ and $\phi^0(t,x)$ are functions defining the wave polarization. The symbol $\imath$ in the exponent denotes the imaginary unit.

Substitution of (3) into (1) and (2) and collection of the terms of order $1/\varepsilon$ yield the equations

$$\Big(-\rho S_t^2 \delta_{il} + C_{ijkl} \frac{\partial S}{\partial x_j} \frac{\partial S}{\partial x_k} \Big) u_l^0 - e_{kij} \frac{\partial S}{\partial x_k} \frac{\partial S}{\partial x_j} \phi^0 = 0, \tag{4}$$

$$-e_{ikl} \frac{\partial S}{\partial x_i} \frac{\partial S}{\partial x_k} u_l^0 - \epsilon_{ij} \frac{\partial S}{\partial x_i} \frac{\partial S}{\partial x_j} \phi^0 = 0 \, . \tag{5}$$

The condition of nontrivial solvability of the system (4) and (5) leads to the eikonal equation

$$\det \left[\frac{1}{\rho} \left(\begin{array}{c|c} C_{ijkl} \frac{\partial S}{\partial x_j} \frac{\partial S}{\partial x_k} & -e_{kij} \frac{\partial S}{\partial x_k} \frac{\partial S}{\partial x_j} \\ \hline -e_{ikl} \frac{\partial S}{\partial x_i} \frac{\partial S}{\partial x_k} & -\epsilon_{ij} \frac{\partial S}{\partial x_i} \frac{\partial S}{\partial x_j} \end{array} \right) - \left(\begin{array}{c|c} \delta_{il} & 0_{3\times 1} \\ \hline 0_{1\times 3} & 0 \end{array} \right) S_t^2 \right] = 0,$$

which can be rewritten as

$$S_t - |\nabla S| c_\alpha \left(\frac{\nabla S}{|\nabla S|} \right) = 0, \tag{6}$$

where $c_\alpha(n)$, $\alpha = 1, 2, 3$, are eigenvalues of the problem

$$\det \left[\frac{1}{\rho} \left(\begin{array}{c|c} C_{ijkl} n_j n_k & -e_{kij} n_k n_j \\ \hline -e_{ikl} n_i n_k & -\epsilon_{ij} n_i n_j \end{array} \right) - \left(\begin{array}{c|c} \delta_{il} & 0_{3\times 1} \\ \hline 0_{1\times 3} & 0 \end{array} \right) c^2 \right] = 0.$$

Here, $n_1, n_2, n_3, |n| = 1$, are components of the normalized wave vector (the direction of propagation). Therefore, for each vector n, there are three types of waves propagating in this direction. Each of them has its own velocity c_α and the corresponding nontrivial solutions, u_l^0 and φ^0, of (4) and (5) defining the wave polarization.

Figure 1a shows the phase velocity surface for $LiTaO_3$ piezoelectric crystals. This surface is obtained as the set of points of the form $c_\alpha(n) \cdot n$, where n belongs to a grid on the surface of the unit sphere. The index α corresponds to a quasi shear wave where the displacements are near orthogonal to the propagation direction. Figure 1b presents the so-called slowness surface consisting of points of the form $c_\alpha^{-1}(n) \cdot n$. It is easy to prove that the slowness surface can be described as $\{p \in R^3 : c_\alpha(p/|p|)|p| = 1\}$. The nonconvexity of the slowness surface shows that the Hamiltonian, $c_\alpha(p/|p|)|p|$, of Eq. (6) is non-convex in p.

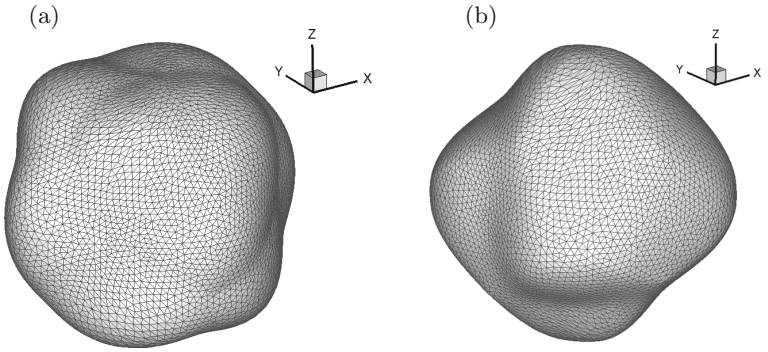


Fig. 1. Characteristic surfaces for lithium tantalate ($LiTaO_3$): (a) Phase velocity surface; (b) Slowness surface. The coordinate frames show the crystalline axes X, Y, and Z.

3 Surface Acoustic Wave Biosensors

Biosensors serve for the measurement of small amounts of biological substances in liquids. Usually a biosensor can be considered as a multi-layered structure (see Fig. 2) whose bottom layer is the ST-cut of piezoelectric α-quartz. Acoustic shear waves are excited here by means of a high-frequency voltage applied to electrodes placed on the ST-cut surface. The waves are transmitted into an isotropic guiding layer deposited on the top of the quartz substrate. The top gold layer is covered by DNA or RNA molecules, aptamers, that are able to specifically bind protein molecules from the contacting liquid. Binding protein molecules results in additional mass loading, which causes a phase shift in the electric signal measured by the output electrodes.

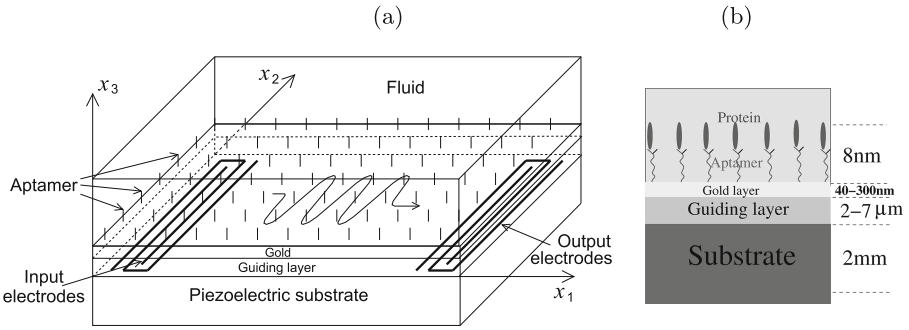


Fig. 2. Structure of acoustic biosensor: (a) Spatial representation; (b) Vertical cross-section with thicknesses of the layers.

A high sensitivity regarding to the added mass is achieved due to the usage of shear horizontally polarized guided waves (Love waves) because of their low

interaction with the contacting fluid. The input and output electrodes are located between the substrate and the guiding layer. To obtain purely shear polarized modes, the direction of the wave propagation is chosen to be orthogonal to the crystalline X-axis (see Fig. 3).

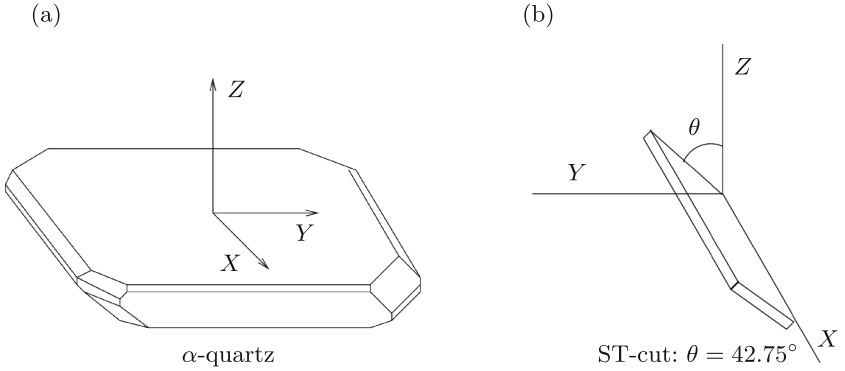


Fig. 3. Piezoelectric α-quartz crystal: (a) Direction of crystalline axes; (b) Orientation of ST-cut.

The next three subsections consider a mathematical model of the biosensor and two methods of numerical investigation of acoustic Love waves including their phase velocity, decay with depth, polarization, etc.

3.1 Mathematical Model of Biosensor

The governing equations for the displacements and the electric potential in the quartz substrate are given by formulae (1) and (2).

The gold layer is conductor so that there is no electric field inside it. The electric field inside the guiding layer is also neglected because of its low dielectric permeability. Therefore, the electric potential vanishes, and the gold and guiding layers are described by the equation of the form

$$\bar{\varrho} u_{i\,tt} - \bar{C}_{ijkl} \frac{\partial^2 u_l}{\partial x_j \partial x_k} = 0\,. \tag{7}$$

In the fluid layer, the Stokes and mass conservation equations hold:

$$\varrho_0 v_{i\,t} - \nu \Delta v_i - (\zeta + \frac{\nu}{3}) \frac{\partial}{\partial x_i} \operatorname{div} v + \frac{\partial}{\partial x_i} \mathcal{P} = 0, \tag{8}$$

$$\gamma \mathcal{P}_t + \frac{\partial}{\partial x_i} v_i = 0,$$

where v_i are components of the velocity, $\mathcal{P}$ is the pressure, ϱ_0 is the fluid density at a reference pressure $\mathcal{P}_0$, ν and ζ are the dynamic and volume viscosities of the fluid, respectively, and $\gamma = \dfrac{1}{\varrho_0} \dfrac{\partial \varrho}{\partial \mathcal{P}}\Big|_{\mathcal{P}_0}$ is the compressibility of the fluid.

A special homogenization technique developed in [3] is used to treat the aptamer-fluid structure. This bristle structure is replaced by an averaged material whose properties are derived as the number of bristles goes to infinity, their thickness tends to zero, and the height remains constant. The resulting new layer whose thickness is equal to the height of the aptamer emulates the aptamer-fluid structure. The governing equation for this layer is given by the relation (see [3,4])

$$\hat{\varrho} u_{i\,tt} - \hat{C}_{ijkl} \frac{\partial^2 u_l}{\partial x_j \partial x_k} - \hat{P}_{ijkl} \frac{\partial^2 u_{lt}}{\partial x_j \partial x_k} = 0, \tag{9}$$

where the term containing the tensor $\hat{P}$ describes the viscous damping coming from the liquid part of the aptamer-fluid structure. The term containing $\hat{C}$ represents elastic stresses. The density $\hat{\varrho}$ is a weighted combination of the densities of the fluid and the aptamer. The tensors $\hat{P}$ and $\hat{C}$ are computed with FE-method using an analytical representation of solutions of the so-called cell equation arising in homogenization theory.

The conditions on the interfaces between the layers are carefully considered in [5,6]. Briefly, the continuity of the displacements and the equilibrium of the normal pressures must hold on the interface between every two neighboring solid layers (the averaged aptamer-fluid layer is considered as solid). Moreover, the electric displacement and the tangent component of the electric field in the substrate must be zero on the interface between the quartz substrate and the guiding layer. The conditions on the interface between the aptamer layer and the fluid include the no-slip assumption and the equilibrium of the pressures.

3.2 Finite Element Modeling

The FE-model extends the above described basic model by accounting for two alternated groups of electrodes (see Fig. 2a) and a damping area around the side and bottom faces to suppress the wave reflection thereon.

The electrodes are typically made of gold. Therefore, they can be accounted for by the linear elasticity equation of type (7).

Accounting for the damping is done by adding the term $-\mathrm{div}(\beta(x)\nabla u_{i\,t})$ to Eqs. (2), (7), and (9), where $\beta(x)$ is a piecewise-linear function which is equal to zero outside of the damping region and grows up to some value $\beta_0 > 0$ towards the side and bottom faces.

The FE-approach provides accurate results because of accounting for the exact parameters of the sensor such as the shape of the electrodes, their position, mass, electro-conductivity properties. This allows us to estimate important characteristics of the biosensor and effects caused by scattering of waves (see [5,6] for simulation results).

The main difficulty of this approach is very high resource-consuming because of a very small wavelength. A large number of finite elements in x_1-direction is required to resolve the wave structure. The number of degrees of freedom lies in the range of 10^6–10^7, which makes impossible, e.g. to compute the phase wave surface with appropriate accuracy.

3.3 Harmonic Analysis (Dispersion Relations)

The approach related to the harmonic analysis is developed in [4] and provides a method for the construction of travelling wave solutions feasible in the biosensor structure under the assumption of its unboundedness in the lateral and downward directions. This assumption is very realistic because real biosensor chips are imbedded up to the surface in very viscous damping media that suppresses the reflection of waves on the side and bottom faces.

The algorithm is described here quite briefly (see [4,7] for more details). It is assumed that all the layers are infinite in x_1 and x_2 directions, the (top) fluid layer and the (bottom) substrate layer are semi-infinite in x_3 direction. The electrodes are not taken into account.

We are looking for solutions describing plain waves propagating in x_1 direction. This means that the displacements in the solid layers, the velocities in the fluid, and the electric potential in the substrate are of the form:

$$u_i(x_1, x_3) = a_i(x_3)\cos(\kappa x_1 - \omega t) + b_i(x_3)\sin(\kappa x_1 - \omega t), \tag{10}$$

$$v_i(x_1, x_3) = c_i(x_3)\cos(\kappa x_1 - \omega t) + d_i(x_3)\sin(\kappa x_1 - \omega t), \tag{11}$$

$$\varphi(x_1, x_3) = f(x_3)\cos(\kappa x_1 - \omega t) + g(x_3)\sin(\kappa x_1 - \omega t), \tag{12}$$

where κ is the wave number, and ω is the circular frequency which is equal to the frequency of the voltage applied to the input electrodes in our case. Substitute (10) and (12) into (1) and (2) for the substrate; (10) into (7) and (9) for non-piezoelectric layers and for the aptamer layer; and (11) into (8) for the fluid layer. Collecting all coefficients of cos and sin yields a system of ordinary linear differential equations for the coefficients a_i, b_i, c_i, d_i, f, g in each layer. Solving these systems for every layer, we obtain the representation of the functions a_i, b_i, c_i, d_i, f, g in the following form (only the expression for the function $a = (a_1, a_2, a_3)$ is given here because the form of the other functions is similar):

$$a(x_3) = \sum_j D^j h^j e^{\lambda^j \kappa x_3}, \tag{13}$$

where D^j are arbitrary coefficients, λ^j and h^j are eigenvalues and eigenvectors of the matrix of the corresponding system of differential equations. For the semi-infinite fluid and substrate layers, only terms decreasing towards x_3 for the fluid and towards $-x_3$ for the substrate, i.e. terms with negative $Re\lambda^j$ for the fluid and positive $Re\lambda^j$ for the substrate, are kept.

Every layer has its own set of coefficients D^j, eigenvalues λ^j, and eigenvectors h^j. To find any particular travelling wave solution in the whole structure we need to determine the coefficients D^j for each layer, which is being done by substituting the expressions of the form (13) for the functions a_i, b_i, c_i, d_i, f, g into (10)–(12) and then the resulting functions u_i, v_i, φ into the interface conditions outlined at the end of Subsect. 3.1 (see [6] for exact description of the interface conditions). Since all of the interface conditions are linear relations, the

computation yields a homogeneous system of linear equations for the unknown coefficients D^j. Denote by $G(\omega, \kappa)$ the matrix of this system. Fix the circular frequency ω and denote the unknown phase velocity by $V = \omega/\kappa$ to consider G as a function of V. The phase velocity is feasible if and only if the system has a nontrivial solution, which is equivalent to the condition $\det \left|\overline{G}^T(V)G(V)\right| = 0$, where $\overline{G}^T(V)$ the conjugate transpose of $G(V)$. The last equation can be easily solved because the computation of the left-hand-side runs very quickly even on an ordinary computer. Usually, there are several roots corresponding to different types of waves propagating with different phase velocities. Concerning the biosensor, the root corresponding to a shear wave, i.e. only $u_2 \neq 0$, is to be chosen.

Figure 4a shows the computed phase velocity contour for surface acoustic waves exited in the biosensor structure using the excitation frequency of 96 MHz. Figure 4b presents the slowness contour scaled by 10^3. It is seen that the slowness contour is not convex (see remark at the end of Sect. 2).

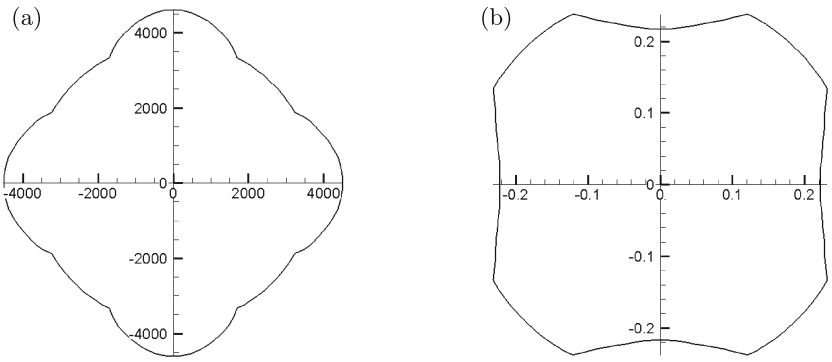


Fig. 4. Characteristic contours for surface acoustic waves: (a) Phase velocity contour; (b) Slowness contour.

4 Description of Wave Propagation

This section addresses the question how to describe the propagation of waves if the velocity surface (contour in the case of surface waves) is known. Let us first recall the classification of surfaces related to the wave propagation. Then, the applicability of Fermat's minimum time principle and the direct usage of eikonal equations will be discussed.

4.1 Characteristic Surfaces

In acoustics, three characteristic surfaces are used to characterize the wave propagation (see [8,9]).

Wave surface. The wave surface (or the group velocity surface) describes the propagation of acoustic energy. This surface is involved in the formulation of minimum time principles, e.g. Fermat's law. The wave surface is the locus of points traced by the energy velocity vector V_e, drawn from a fixed point O, as the propagation direction varies. The propagation direction n (the normalized wave vector) is orthogonal to the wave surface (see Fig. 5a). It should be noticed that V_e is as a rule not collinear to the wave vector in the case of anisotropic media.

Phase velocity surface. The phase velocity surface (see Figs. 1a and 4a) describes the propagation of wave fronts. It defines the Hamiltonian of the eikonal equation. The phase velocity surface is obtained from the wave surface by projecting the vector V_e onto the wave propagation direction n (see Fig. 5b) so that the phase velocity vector V is given by $V = (V_e \cdot n)\, n$. It should be noticed that the phase velocity surface can be constructed independently on V_e, e.g. as shown in Sect. 2 and Subsect. 3.3, and the wave surface can then be defined through the phase velocity surface.

Slowness surface. The slowness surface (see Figs. 1b and 4b) indicates the local convexity/concavity properties of the Hamiltonian of the eikonal equation. The slowness surface is related to the phase velocity surface by the inversion through the origin (see Fig. 5c), i.e. $m = n/|V|$. The energy velocity, V_e, is normal to the slowness surface at all points. Local concavities on the slowness surface can cause formation of cusps ("swallow tails") on the wave surface as it is shown in Fig. 5c: The arc (acb) is mapped into a "swallow tail" on the wave surface. This points out to the intersection of characteristics of the eikonal equation.

4.2 Fermat's Principle

The Fermat principle describes how a ray, trajectory orthogonal to the wave front at all time instants, propagates from point A to point B. The principle says that the propagation time should be minimal. To express this, consider the minimization problem

$$T = \int_A^B dt = \int_A^B \frac{ds}{|V_e|} \to \min,$$

where the integrals are computed along rays. Let $x(\tau)$ be the parametrization of rays. Accounting for the relation $ds = |\dot{x}|d\tau$ yields

$$T = \int_{\tau_0}^{\tau_1} \frac{|\dot{x}|}{|V_e(x, \dot{x}/|\dot{x}|)|} d\tau =: \int_{\tau_0}^{\tau_1} L(x, \dot{x}) d\tau \to \min .$$

Thus, feasible rays are solutions of the Euler equation

$$L_x - \frac{d}{d\tau} L_{\dot{x}} = 0 .$$

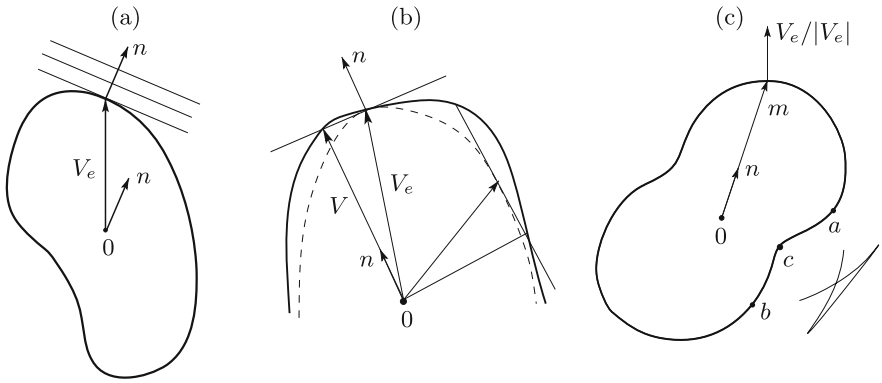


Fig. 5. Schematic explanation to characteristic surfaces: (a) Wave surface; (b) Phase velocity surface; (c) Slowness surface.

This approach works well if the wave velocity $V_e(x, n)$ is well-defined for all directions n. It holds if the slowness surface is convex, which is as a rule violated in the case of anisotropic media. The next subsection discusses the method of direct solving eikonal equations.

4.3 Eikonal and Hamilton-Jacobi Equations

Let $c(x, n)$ be the phase velocity depending on the spatial position x and the propagation direction n, $|n| = 1$, Let $S(t, x)$ be the phase function that shows the phase of the wave at the time instant t and at the point x. According to the results of Sect. 2, cf. Eq. (6), the eikonal equation reads

$$S_t - |\nabla S| c\left(x, \frac{\nabla S}{|\nabla S|}\right) = 0. \tag{14}$$

If the Hamiltonian $c(x, p/|p|)|p|$ is convex in p, the method of characteristics can be used for solving Eq. (14) in the case of convex wave emitter. Under these conditions, the characteristics representing the rays do not intersect each other. If the convexity property is violated, Eq. (14) may not have classical solutions. Nevertheless, it is always uniquely solvable in the sense of viscous solutions (see [10,11]). Moreover, a unique viscous solution of (14) is the valid phase function. Thus, we arrive at the idea to use numerical methods of finding viscosity solutions of Hamilton-Jacobi equations. It should be noticed that common Lax-Friedrichs methods (see e.g. [12]) are not applicable in this case because they smooth solutions very strong. On the other hand, the authors have developed numerical methods that do not contain any smoothing (see e.g. [13,14]). These methods assume that the Hamilton-Jacobi equations arise from conflict control problems (differential games). Therefore, their application requires solving the following problem: Given an eikonal equation, it is required to construct a differential game whose Hamiltonian coincides (up to the sign) with that of the

eikonal equation. The next section shows how to do that using an unification technique proposed in [15].

5 Usage of Differential Games

Assume that a set $M \subset R^d$, $d = 2$ or 3, represents the shape of the acoustic wave emitter. For example, M is a ball in the case of bulk crystal, and M is the two dimensional area of the input electrodes in the case of biosensor. Let the game dynamics be described by the following system of ordinary differential equations:

$$\dot{x} = f(x, u, v), \quad x \in R^d, \ t \in (-\infty, 0], \ u \in P \subset R^a, \ v \in Q \subset R^b, \tag{15}$$

where u and v are control parameters of the first and second players, respectively. Introduce the signed distance, σ, to the set M as follows: $\sigma(x) = \mathrm{dist}(x, M)$ if $x \notin M$, and $\sigma(x) = -\mathrm{dist}(x, R^d \backslash M)$ if $x \in M$. Consider the objective functional, γ, defined on the trajectories of (15) as follows:

$$\gamma(x(\cdot)) = \min_{\tau \in [t,0]} \sigma(x(\tau)). \tag{16}$$

The game is formalized using the concept of feedback strategies (see [16]). The value function is defined by the relation

$$\Psi(t, x) = \max_{\mathcal{V}} \min_{x(\cdot) \in X(t,x,\mathcal{V})} \gamma(x(\cdot)),$$

where $\mathcal{V}$ is a feedback strategy of the second player, and the set $X(t, x, \mathcal{V})$ expresses the actions of the first player. This set consists of all limits of Euler trajectories of (15) which are obtained when the second player chooses $v \equiv \mathcal{V}(t_i, x(t_i))$ on each interval $[t_i, t_{i+1})$ of partitions of $[t, 0]$, and the first player uses admissible controls $u(\xi)$, $\xi \in [t, 0]$. In doing that, all possible partitions whose diameter tends to zero and all admissible controls of the first player are exhausted. All Euler trajectories start at t from the initial state x.

The value function is locally bounded and Lipschitzian (see e.g. [11]).

Define the Hamiltonian

$$H(x, p) = \max_{v \in Q} \min_{u \in P} \langle p, f(x, u, v) \rangle, \quad p \in R^d, \tag{17}$$

and consider the Hamilton-Jacobi-Bellman-Isaacs equation

$$\Psi_t + H(x, \Psi_x) = 0, \quad \Psi(0, x) = \sigma(x). \tag{18}$$

It is proven in [13] that the value function of the game (15) with the objective functional (16) is a viscosity solution of (18). Therefore, the following proposition holds:

Proposition 1. *Let* $c\big(x, p/|p|\big)\,|p|$ *be the Hamiltonian of the eikonal equation. If*

$$H(x,p) = -c\big(x, p/|p|\big)\,|p|, \quad p \in R^d,$$

then the wave front at any time instant $t \geq 0$ *is given by the relation*

$$\{x : \Psi(\text{-}t, x) = 0\}.$$

This proposition opens the way to use numerical methods of finding viscosity solutions of Hamilton-Jacobi equations. The only question consists in constructing an appropriate differential game whose Hamiltonian satisfies the condition of Proposition 1. The next subsection discusses this task.

5.1 Unification

Denote $E(x,p) = -c\big(x, p/|p|\big)\,|p|$, $p \in R^d$. To find a differential game whose Hamiltonian coincides with E, the technique of unification (see [15]) can be used. Consider the following conflict control system

$$\dot{x} = E(x,p)p + q, \quad x, p, q \in R^d, \quad |p| = 1, \quad |q| = \lambda, \quad \langle p, q\rangle \geq 0. \tag{19}$$

Here, q is the control parameter of the first player who strives to minimize the objective functional (16), whereas p is the control parameter of the second player who maximizes the objective functional. The parameter λ is a constant which is greater than the Lipschitz constant of the function E in p.

Proposition 2. *If* $|E(x,p_1) - E(x,p_2)| < \lambda|p_1 - p_2|$, $p_1, p_2 \in R^d$, $|p_1| = 1$, $|p_2| = 1$, *then the Hamiltonian of the game (19) satisfies the relation*

$$H_{(19)}(x,s) := \max_{|p|=1} \min_{\substack{|q|\,=\,\lambda,\\ \langle p,\,q\rangle\,\geq\,0}} \langle E(x,p)p + q, s\rangle = E(x,s),$$

and, therefore, (19) is the required differential game.

It should be noticed that the proof of this proposition essentially uses the positive homogeneity of the function $p \to E(\cdot, p)$. Therefore, the unification procedure cannot be applied in the case of absence of positive homogeneity.

Assume now that $|p| = 1$, then $E(x,p)p = -c(x,p)p$. Taking into account that the points $c(x,p)p$ exhaust the phase velocity surface $V_{\text{surf}}(x)$ when varying p, the game (19) can be rewritten as

$$\dot{x} = -p + q, \quad p \in V_{\text{surf}}(x), \quad |q| = \lambda, \quad \langle p, q\rangle \geq 0.$$

Assuming that only the velocity magnitude depends on the spatial position yields the game

$$\dot{x} = -a(x)p + q, \quad p \in V_{\text{surf}}, \quad |q| = \lambda, \quad \langle p, q\rangle \geq 0, \tag{20}$$

where the phase velocity surface V_{surf} does not depend on x. Notice that the coefficient $a(x)$ is necessary to take into account the damping area of the biosensor, the phase velocity is strongly reduced there. Therefore, $a(x) \equiv 1$ outside the damping area, and $a(x) \to 0$ towards the outer boundary.

Numerical methods developed by the authors (see [13,14]) provide an effective tool for finding the value function of the game (20) with the objective functional (16). The next section uses these methods and demonstrates a good reconstruction of wave patterns.

6 Simulation of Wave Propagation Using the Unified Differential Game

Now, the differential game (20) with the objective functional (16) is used for the computation of wave fronts. First, consider waves in $LiTaO_3$ piezoelectric crystals. Assume that the wave emitter, the set M, is a ball of radius 0.1, and $a(x) \equiv 1$, i.e. there is no damping area. The phase velocity surface is shown in Fig. 1a. Thus, all data required for the formulation of the differential game (16) and (20) are available. Application of a finite difference upwind scheme described in [13] yields an approximation, $\tilde{\Psi}$, of the value function of the game. According to Proposition 1, the wave front at any time instant $t \geq 0$ is given by the relation $\{x \in R^3 : \tilde{\Psi}(-t, x) = 0\}$. Figure 6 shows the wave front at $t = 0.25, 1$, and 2 ms.

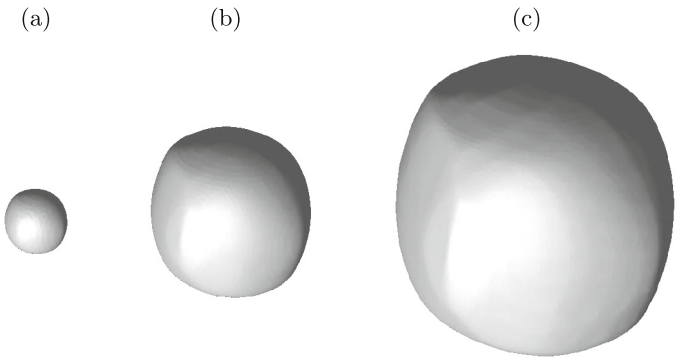


Fig. 6. Wave fronts in lithium tantalate $LiTaO_3$ piezoelectric crystal (quasi shear wave): (a) $t = 0.25$ ms; (b) $t = 1$ ms; (c) $t = 2$ ms.

Consider now wave fronts for surface shear waves in the biosensor structure (see Sect. 3). The phase velocity contour is shown in Fig. 4a. Figure 7 shows the position of the wave front for a time sequence with a small sample time. The wave emitter and the damping area are easily recognizable in this figure.

Figures 8 and 9 show the case of two wave emitters. Figure 10 shows the wave propagation in the presence of a hole. The hole is interpreted as an obstacle such that the trajectories of the game (20) can not penetrate therein. This case is numerically processed using a method for finding value functions in games with state constraints (see [14]).

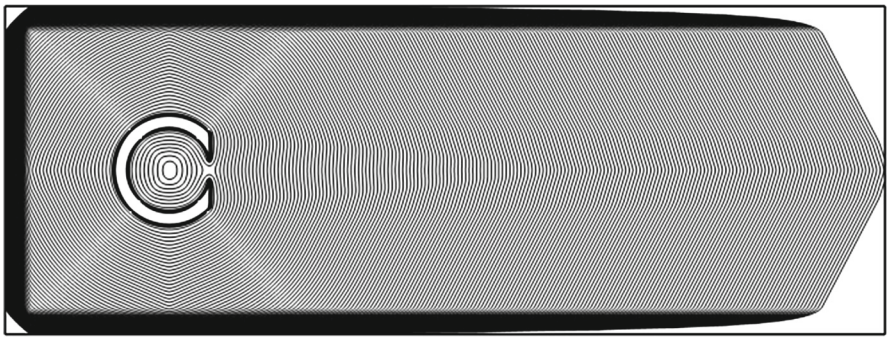


Fig. 7. Surface wave fronts in the biosensor structure. The wave emitter is an unclosed ring.

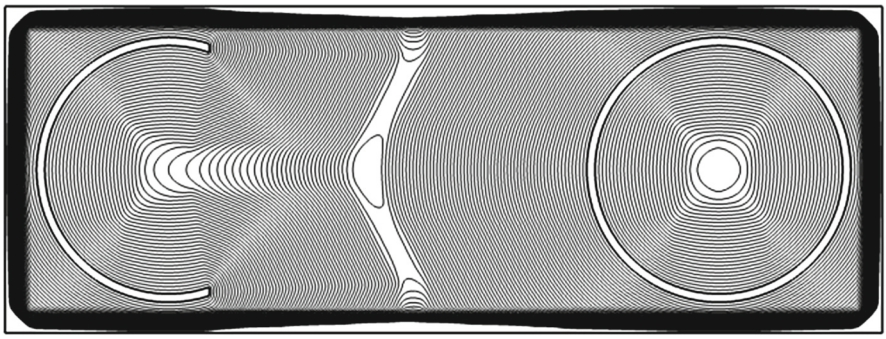


Fig. 8. Surface wve fronts in the biosensor structure with two emitters: a half-ring and a ring.

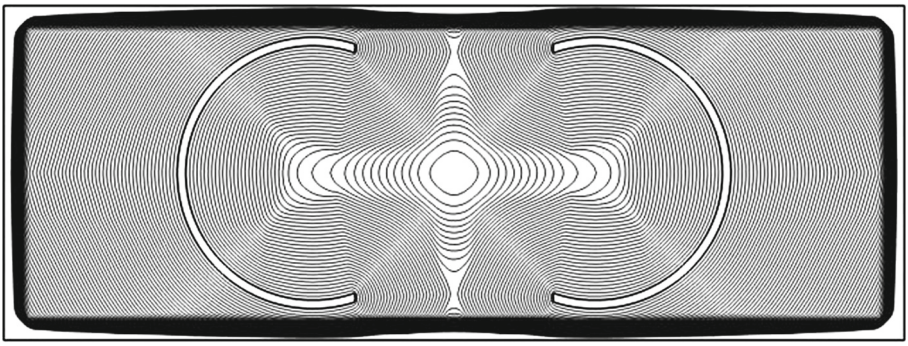


Fig. 9. Surface wave fronts in the biosensor structure with two half-rings as the emitters.

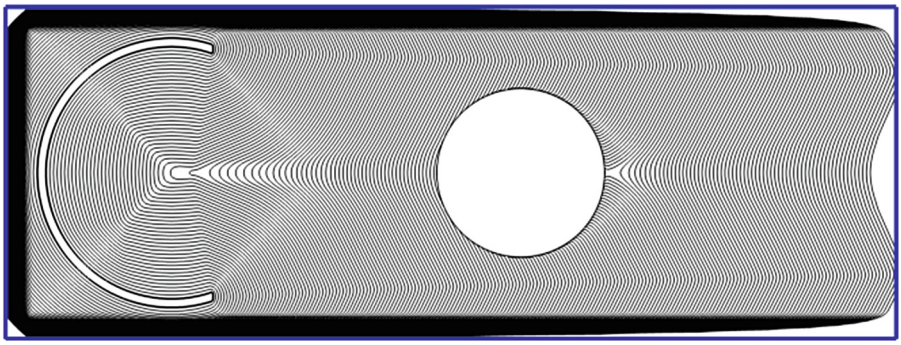


Fig. 10. Surface wave fronts in the biosensor structure in the presence of a round hole. The wave emitter is a half-ring.

7 Conclusion

The technique presented in this paper is also appropriate for the numerical treatment of arbitrary Hamilton-Jacobi equations with positive homogeneous Hamiltonians. As it was seen, the unification procedure described in Sect. 5.1 does not use any specific features of the Hamiltonian with the exception of the positive homogeneity which is necessary for the proof of Proposition 2. This opens new possibilities of investigation of physical processes related to optimality principles involving non-convex Lagrangians and Hamiltonians.

References

1. Melikyan, A.A.: Generalized Solutions of First Order PDEs. Birkhäuser, Boston (1998)
2. Barles, G.: Remarks on a flame propagation model. Technical report 464, INRIA (1985)
3. Hoffmann, K.-H., Botkin, N.D., Starovoitov, V.N.: Homogenization of interfaces between rapidly oscillating fine elastic structures and fluids. SIAM J. Appl. Math. **65**(3), 983–1005 (2005)
4. Botkin, N.D., Hoffmann, K.-H., Pykhteev, O.A., Turova, V.L.: Dispersion relations for acoustic waves in heterogeneous multi-layered structures contacting with fluids. J. Franklin Inst. **34**(5), 520–534 (2007)
5. Botkin, N.D., Turova, V.L.: Mathematical models of a biosensor. Appl. Math. Model. **28**(6), 573–589 (2004)
6. Botkin, N.D., Hoffmann, K.-H., Pykhteev, O.A., Starovoitova, B.N., Turova, V.L.: Two complementary approaches in modelling a biosensor. In: Hamza. M.N. (ed.) 15th IASTED International Conference on Aplied Simulation and Modelling, pp. 525–530. ACTA Press, Anaheim-Calgary-Zurich (2006)
7. Botkin, N.D., Hoffmann, K.-H., Pykhteev, O.A., Turova, V.L.: Numerical computation of dispersion relations for multi-layered anisotropic structures. In: Technical Proceedings of 2004 Nanotechnology Conference and Trade Show, vol. 2, pp. 411–414. NSTI, Boston (2004)

8. Auld, B.A.: Acoustic Fields and Waves in Solids. I. Krieger Publishing Company, Malabar (1972)
9. Royer, D., Dieulesaint, E.: Elastic Waves in Solids I: Free and Guided Propagation. Springer, Heidelberg (2000)
10. Crandall, M.G., Lions, P.L.: Viscosity solutions of Hamilton-Jacobi equations. Trans. Amer. Math. Soc. **277**, 1–47 (1983)
11. Subbotin, A.I.: Generalized Solutions of First Order PDEs: The Dynamical Optimization Perspective. Birkhäuser, Boston (1995)
12. Mitchell, I.: Application of level set methods to control and reachability problems in continuous and hybrid systems. Ph.D. Thesis, Stanford University (2002)
13. Botkin, N.D., Hoffmann, K.-H., Turova, V.L.: Stable numerical schemes for solving Hamilton-Jacobi-Bellman-Isaacs equations. SIAM J. Sci. Comput. **33**(2), 992–1007 (2011)
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. Krasovskii, N.N.: On the problem of the unification of differential games. Dokl. Akad. Nauk SSSR **226**(6), 1260–1263 (1976)
16. Krasovskii, N.N., Subbotin, A.I.: Game-Theoretical Control Problems. Springer, New York (1988)

Efficient Cardinality/Mean-Variance Portfolios

R. Pedro Brito[1] and Luís N. Vicente[2(✉)]

[1] Department of Mathematics, University of Coimbra,
3001-454 Coimbra, Portugal
rpedro.brito@gmail.com
[2] CMUC, Department of Mathematics, University of Coimbra,
3001-454 Coimbra, Portugal
lnv@mat.uc.pt

Abstract. We propose a novel approach to handle cardinality in portfolio selection, by means of a biobjective cardinality/mean-variance problem, allowing the investor to analyze the efficient tradeoff between return-risk and number of active positions. Recent progress in multiobjective optimization without derivatives allow us to robustly compute (in-sample) the whole cardinality/mean-variance efficient frontier, for a variety of data sets and mean-variance models. Our results show that a significant number of efficient cardinality/mean-variance portfolios can overcome (out-of-sample) the naive strategy, while keeping transaction costs relatively low.

Keywords: Portfolio selection · Cardinality · Sparse portfolios · Multiobjective optimization · Efficient frontier · Derivative-free optimization

1 Introduction

One knows since the pioneer work of Markowitz [19] that a rational investor has typically two goals in mind: to maximize the portfolio return (given, e.g., by the portfolio expected return) and to minimize the portfolio risk (described, e.g., by the portfolio variance). Traditionally, the Markowitz mean-variance optimization model is taken as a quadratic program (QP), intended to minimize the portfolio risk (variance) for a given level of expected return, over a set of feasible portfolios. By varying the level of expected return, the Markowitz model determines the so-called efficient frontier, as the set of nondominated portfolios regarding the two goals (variance and mean of the return). The rational investor can thus make choices, by analyzing the tradeoff between expected return and variability of the investment, over a set of appropriate portfolios.

Several modifications to the classical Markowitz model or alternative methodologies have since then been proposed. One resulting from a simple observation was suggested in an article by DeMiguel, Garlappi, and Uppal [14]. These authors

Luís Vicente: Support for this research was provided by FCT under the grant PTDC/MAT/098214/2008.

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 52–73, 2014.
DOI: 10.1007/978-3-662-45504-3_6

analyzed a number of methodologies inspired on the classic model of Markowitz and showed that none were able to significantly and consistently overcome the naive strategy, that is to say, the one in which the available investor's wealth is divided equally among the available securities. One possible explanation is related to the ill conditioning of the objective function of the Markowitz model (given by the variance of the return).

One of the important issues to consider in portfolio selection is how to handle transaction costs. There are well known modifications that can be made in the Markowitz model to incorporate transaction costs, such as to bound the turnover, which basically amount to further linear constraints in the QP. A recent technique to keep transaction costs low consists of selecting sparse portfolios, i.e., portfolios with few active positions, by imposing a cardinality constraint. Such a constraint, however, changes the classical QP into a MIQP (mixed-integer quadratic programming), which can no longer be solved in polynomial time.

In this paper, we suggest an alternative approach to the cardinality constrained Markowitz mean-variance optimization model, reformulating it directly as a biobjective problem, allowing the investor to analyze the tradeoff between cardinality and mean-variance, in a general scenario where short-selling is permitted. Such an approach allows us to find the set of nondominated points of biobjective problems in which an objective is smooth and combines mean and variance and the other is nonsmooth (the cardinality or ℓ_0 norm of the vector of portfolio positions). The mean-variance objective function can take a number of forms. A parameter free possibility is given by profit per unity of risk (a nonlinear function obtained by dividing the expected return by its variance).

Given the lack of derivatives of the cardinality function, we decided then to apply a directional derivative-free algorithm for the solution of the biobjective optimization problem. Such methods do not require derivatives, although their convergence results typically assume some weak form of smoothness such as Lipschitz continuity. Direct multisearch is a derivative-free multiobjective methodology for which one can show some type of convergence in the discontinuous case. More importantly, it exhibited excellent numerical performance on a comparison to a number of other multiobjective optimization solvers. We applied direct multisearch to determine (in-sample) the set of efficient or nondominated cardinality/mean-variance portfolios.

To illustrate our approach, we gathered several data sets from the FTSE 100 index (for returns of single securities) and from the Fama/French benchmark collection (for returns of portfolios), computed the efficient cardinality/mean-variance portfolios using (in-sample) optimization, and measured their out-of-sample performance using a rolling-sample approach. We found that a large number of sparse portfolios for the FTSE 100 data sets, among the efficient cardinality/mean-variance ones, consistently overcome the naive strategy in terms of out-of-sample performance measured by the Sharpe ratio. This effect is also clearly visible for the FF data sets, where the performance of a large portion of the cardinality/mean-variance efficient frontier outperforms, in most of the instances, the naive strategy. The transactions costs are shown to be

relatively low for all efficient cardinality/mean-variance portfolios, with a moderate increase with cardinality.

The organization of our paper is as follows. In the next section, we formulate the classical Markowitz model for portfolio selection, describe the naive strategy, and formulate the problem with cardinality constraint. In Sect. 3, we reformulate the cardinality constrained Markowitz mean-variance optimization model as a biobjective problem for application of multiobjective optimization. In Sect. 4, we present the empirical results. Finally, in Sect. 5 we summarize our findings and discuss future research.

2 Portfolio Selection Models

2.1 The Classical Markowitz Mean-Variance Model

Portfolios consist of securities (shares or bonds, for example, or classes or indices of the same). Suppose the investor has a certain wealth to invest in a set of N securities. The return of each security i is described by a random variable R_i, whose average can be computed (from estimation based on historical data). Let $\mu_i = E(R_i)$, $i = 1, \ldots, N$, denote the expected returns of the securities. Let also w_i, $i = 1, \ldots, N$, represent the proportions of the total investment to allocate in the individual securities. The portfolio return is assumed linear in $w_1, \ldots, w_N$, and thus the portfolio expected return can be written as

$$E(R) \;=\; E(w_i R_1 + \cdots + w_N R_N) \;=\; w_1\mu_1 + \cdots + w_N\mu_N \;=\; \mu^\top w$$

with

$$\mu \;=\; (\mu_1, \ldots, \mu_N)^\top \quad \text{and} \quad w \;=\; (w_1, \ldots, w_N)^\top.$$

The portfolio variance, in turn, is calculated by

$$V(R) \;=\; E\Bigg(\bigg[\sum_{i=1}^{N} w_i R_i - E\bigg(\sum_{i=1}^{N} w_i R_i\bigg)\bigg]^2\Bigg).$$

So,

$$V(R) \;=\; \sum_{i=1}^{N}\sum_{j=1}^{N} E[(R_i - \mu_i)(R_j - \mu_j)]w_i w_j.$$

Representing each entry i, j of the covariance matrix Q by

$$\sigma_{ij} = E[(R_i - \mu_i)(R_j - \mu_j)],$$

one has

$$V(R) \;=\; w^\top Q w,$$

where Q is symmetric and positive semi-definite (and typically assumed positive definite). As said before, a portfolio is defined by an $N \times 1$ vector w of weights

representing the proportion of the total funds invested in the N securities. This vector of weights is thus required to satisfy the constraint

$$\sum_{i=1}^{N} w_i \;=\; e^\top w \;=\; 1,$$

where e is the $N \times 1$ vector of entries equal to 1. Lower bounds on the variables, of the form $w_i \geq 0$, $i = 1, \ldots, n$, can be also considered if short selling is undesirable. In general, we will say that $L_i \leq w_i \leq U_i$, $i = 1, \ldots, N$, for given lower L_i and upper U_i bounds on the variables.

Markowitz's model [19,20] is based on the formulation of a mean-variance optimization problem. By solving this problem, we identify a portfolio of minimum variance among all which provide an expected return not below a certain target value r. The aim is thus to minimize the risk from a given level of return. The formulation of this problem can be described as:

$$\begin{array}{ll} \min\limits_{w \in \mathbb{R}^N} & w^\top Q w \\ \text{subject to} & \mu^\top w \geq r, \\ & e^\top w = 1, \\ & L_i \leq w_i \leq U_i, \quad i = 1, \ldots, N. \end{array} \tag{1}$$

Problem (1) is a convex quadratic programming problem (QP), for which the first order necessary conditions are also sufficient for (global) optimality. See [12,21] for a survey of portfolio optimization. The classical Markowitz mean-variance model can be seen as way of solving the biobjective problem which consists of simultaneously minimizing the portfolio risk (variance) and maximizing the portfolio profit (expected return)

$$\begin{array}{ll} \min\limits_{w \in \mathbb{R}^N} & w^\top Q w \\ \max\limits_{w \in \mathbb{R}^N} & \mu^\top w \\ \text{subject to} & e^\top w = 1, \\ & L_i \leq w_i \leq U_i, \quad i = 1, \ldots, N. \end{array} \tag{2}$$

In fact, it is easy to prove that a solution of (1) is nondominated, efficient or Pareto optimal for (2). Efficient portfolios are thus the ones which have the minimum variance among all that provide at least a certain expected return, or, alternatively, those that have the maximal expected return among all up to a certain variance. The efficient frontier (or Pareto front) is typically represented as a 2-dimensional curve, where the axes correspond to the expected return and the standard deviation of the return of an efficient portfolio.

2.2 The Naive Strategy 1/*N*

The naive strategy is the one in which the available investor's wealth is divided equally among the securities available

$$w_i = \frac{1}{N}, \quad i = 1, \ldots, N.$$

This strategy has diversification as its main goal, it does not involve optimization, and it completely ignores the data.

Although a number of theoretical models have been developed in the last years, many investors pursuing diversification revert to the use of the naive strategy to allocate their wealth (see [4]). DeMiguel, Garlappi, and Uppal [14] evaluated fourteen models across seven empirical data sets and showed that none is consistently better than the naive strategy. A possible explanation for this phenomenon lies on the fact that the naive strategy does not involve estimation and promotes 'optimal' diversification. The naive strategy is therefore an excellent benchmarking strategy.

2.3 The Cardinality Constrained Markowitz Mean-Variance Model

Since the appearance of the classical Markowitz mean-variance model, a number of methodologies have been proposed to render it more realistic. The classical Markowitz model assumes a perfect market without transaction costs or taxes, but such costs are an important issue to consider as far as the portfolio selection is concerned, especially for small investors. Recently, it has been studied the addition of a constraint that sets an upper bound on the number of active positions taken in the portfolio, in an attempt to improve performance and reduce transactions costs. Such a cardinality constraint is defined by limiting $\text{card}(x) = |\{i \in \{1, ..., N\} : x_i \neq 0\}|$ and leads to cardinality constrained portfolio selection problems. In particular, the cardinality constrained Markowitz mean-variance optimization problem has the form:

$$\begin{aligned} \min_{w \in \mathbb{R}^N} \quad & w^\top Q w \\ \text{subject to } & \mu^\top w \geq r, \\ & \text{card}(w) \leq K, \\ & e^\top w = 1, \\ & L_i \leq w_i \leq U_i, \quad i = 1, \ldots, N, \end{aligned} \tag{3}$$

where $K \in \{1, \ldots, N\}$. Although $\text{card}(x)$ is not a norm, it is frequently called the ℓ_0 norm in the literature, $\|x\|_0 = \text{card}(x)$. By introducing binary variables, one can rewrite the problem as a mixed-integer quadratic programming (MIQP) problem:

$$\begin{aligned} \min_{w,y \in \mathbb{R}^N} \quad & w^\top Q w \\ \text{subject to } & \mu^\top w \geq r, \\ & e^\top y \leq K, \\ & e^\top w = 1, \\ & L_i y_i \leq w_i \leq U_i y_i, \quad i = 1, \ldots, N, \\ & y_i \in \{0, 1\}, \quad i = 1, \ldots, N. \end{aligned} \tag{4}$$

However such MIQPs are known to be hard combinatorial problems. The number of sparsity patterns in w (i.e., number of different possibilities of having K nonzeros entries) is $\binom{N}{K} = N!/[(N-K)!K!]$. Although there are exact algorithms for the solution of MIQPs (see [5–7,25]), many researchers and portfolio managers

prefer to use heuristics approaches (see [3,9,11,15,17,26]). Some of these heuristics vary among evolutionary algorithms, tabu search, and simulated annealing (see [15,26]).

Promotion of sparsity is also used in the field of signal and imaging processing, where a new technique called compressed sensing has been intensively studied in the recent years. Essentially one aims at recovering a desired signal or image with the least possible amount of basis components. The major developments in compressed sensing have been achieved by replacing the ℓ_0 norm by the ℓ_1 one, the latter being a convex relation of the former and known to also promote sparsity. The use of the ℓ_1 norm leads to recovering optimization problems solvable in polynomial time (in most of the cases equivalent to linear programs), and a number of sparse optimization techniques have been developed for the numerical solution of such problems. These ideas have already been used in portfolio selection primarily to promote regularization of ill conditioning (of the estimation of data or of the variance of the return itself). DeMiguel et al. [13] constrained the Markowitz classical model by imposing a bound on the ℓ_1 norm of the vector of portfolio positions, among other possibilities. Brodie et al. [8] focus on a modification to the Markowitz mean-variance classical model by the incorporation of a term involving a multiple of the ℓ_1 norm of the vector of portfolio positions. Inspired by sparse reconstruction (see, for instance, [7]), they also proposed an heuristic for the solution of the problem.

3 The Cardinality/Mean-Variance Biobjective Model

Although the cardinality constrained Markowitz mean-variance model described in (3) provides an alternative to the classical Markowitz model in the sense of realistically limiting the number of active positions in a portfolio, it is dependent on the parameter K, the maximum number of such positions. Thus, one has to vary K to obtain various levels of cardinality or sparsity, and for each value of K solve an MIQP of the form (4).

The alternative suggested in this paper is to consider the cardinality function as an objective function itself. At a first glance, one could see the problem as a triobjective optimization problem by minimizing the variance of the return, maximizing the expected return, and minimizing the cardinality over the set of feasible portfolios. Such a framework was taken into account in the studies [1,2,10,18]. However, these authors did not investigate the effects of cardinality constraints on portfolio models in terms of out-of-sample performance, a subject still poorly analyzed in the literature. On the other hand, investors may find it useful to directly analyze the tradeoff between cardinality and mean-variance. A parameter-free possibility is to consider a Sharpe ratio type objective function, by maximizing expected return per variance and minimizing the cardinality, over the set of feasible portfolios. In this case, the cardinality/mean-variance biobjective optimization problem is posed as

$$\begin{array}{ll} \min\limits_{w\in\mathbb{R}^N} & -\frac{\mu^\top w}{w^\top Q w} \\ \min\limits_{w\in\mathbb{R}^N} & \text{card}(w) \\ \text{subject to } & e^\top w = 1, \\ & L_i \leq w_i \leq U_i, \quad i = 1, \ldots, N. \end{array} \tag{5}$$

By solving (5), we identify a cardinality/mean-variance efficient frontier. A portfolio in this frontier is such that there exists no other feasible one which simultaneously presents a lower cardinality and a lower mean-variance measure. Given such an efficient frontier and a mean-variance target, an investor may directly find the answers to the questions of what is the optimal (lowest) cardinality level that can be chosen and what are the portfolios leading to such a cardinality level. Problem (5) has two objective functions and linear constraints. The first objective $f_1(w) = -\mu^\top w/w^\top Q w$ is nonlinear but smooth. However, the second objective function $f_2(w) = \text{card}(w) = |\{i \in \{1, ..., N\} : w_i \neq 0\}|$ is piecewise linear discontinuous, consequently nonlinear and nonsmooth. We have thus decided to solve the biobjective optimization problem (5) using a derivative-free solver, based on direct multisearch.

4 Empirical Performance of Efficient Cardinality/Mean-Variance Portfolios

Now we report a number of experiments made to numerically determine and assess the efficient cardinality/mean-variance frontier. We applied direct multisearch to determine the Pareto front or efficient frontier of the biobjective optimization problem (5) (according to Appendix A). We tested three data sets collected from the FTSE 100 index and three others from the Fama/French benchmark collection (see Subsect. 4.1). The efficient frontiers obtained by the initial in-sample optimization are given in Subsect. 4.2.

The out-of-sample performance of the cardinality/mean-variance efficient portfolios, measured by a rolling-sample approach, is described in Subsect. 4.3. In Subsect. 4.4 we measure the out-of-sample performance by the Sharpe ratio, in Subsubsect. 4.5 we report the proportional transaction costs, and in Subsect. 4.6 we measure the out-of-sample performance by the Sharpe ratio of returns net of transaction costs, all of this for each cardinality/mean-variance efficient portfolio. To better assess the robustness of our results, we also considered, using the FTSE 100 data, a sample including the financial crisis years 2008–2010, and the corresponding results are reported in Subsect. 4.7. The section is ended with a discussion of the overall obtained results.

4.1 Data Sets

For the first three data sets we collected daily data for securities from the FTSE 100 index, from 01/2003 to 12/2007 (five years). Such data is public and available from the site http://www.bolsapt.com. The three data sets are referred

to as DTS1, DTS2, and DTS3, and are formed by 12, 24, and 48 securities, respectively. The composition of these data sets is given in Table 1. We used the daily continuous returns for the in-sample optimization (estimation of Q and μ) and the daily discrete returns for the out-of-sample analysis. We also included in our experiments three data sets from the Fama/French benchmark collection (FF10, FF17, and FF48, with cardinalities 10, 17, and 48), using the monthly returns from 07/1971 to 06/2011 (forty years) given there for a number of industry security sectors. More information on these security sectors (or portfolios of securities) can be found in http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

Table 1. Composition of the three data sets from the FTSE 100 index. In brackets we indicate the data set to which each security belongs to.

SECURITIES	
3 I GROUP (1,2,3)	JOHNSON MATTHEY P (3)
AMEC (1,2,3)	LEGAL &GENERAL (3)
ANGLO AMERICAN (1,2,3)	LLOIDS BANKING GR (3)
ANTOFAGASTA (1,2,3)	LONMIN (3)
ASSOCIAT BRIT FOO (1,2,3)	MARKS &SPENCER (3)
ASTRAZENECA (1,2,3)	MORRINSON SUPERMKT (3)
AVIVA (1,2,3)	NEXT (3)
B SKY B GROUP (1,2,3)	OLD MUTUAL (3)
BAE SYSTEMS (1,2,3)	PEARSON (3)
BARCLAYS (1,2,3)	PRUDENTIAL (3)
BG GROUP (1,2,3)	REED ELSEVIER PLC (3)
BHP BILLITON (1,2,3)	RENTOKIL INITIAL (3)
BP (2,3)	REXAM (3)
BRIT AMER TOBACCO (2,3)	RIO TINTO (3)
BRIT LAND CO REIT (2,3)	ROYAL BK SCOTL GR (3)
BRITISH AIRWAYS (2,3)	RSA INSUR GRP (3)
CAB &WIRE WRLD (2,3)	SABMILLER (3)
CAPITA GRP (2,3)	SAGE GRP (3)
COBHAM (2,3)	SAINSBURY (3)
DIAGEO (2,3)	SCHRODERS (3)
HAMMERSON REIT (2,3)	SEVERN TRENT (3)
IMPERIAL TOBACCO (2,3)	SHIRE (3)
INTERNATIONAL POW (2,3)	UNITED UTILITIES (3)
INVENSYS (2,3)	VODAFONE GRP (3)

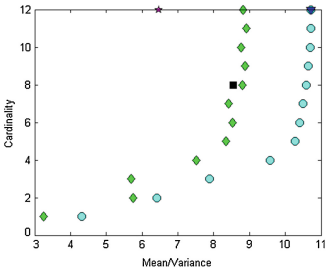


Fig. 1. Efficient frontier of the biobjective cardinality/mean-variance problem for DTS1. ★ Naive ▼ Markowitz mean per variance ■ Markowitz minimum variance ● cardinality/mean-variance ♦ cardinality constrained minimum variance (Color figure online)

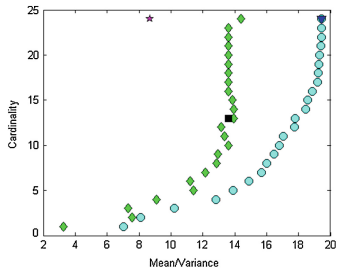


Fig. 2. Efficient frontier of the biobjective cardinality/mean-variance problem for DTS2. See the caption of Fig. 1 for an explanation of the various symbols.

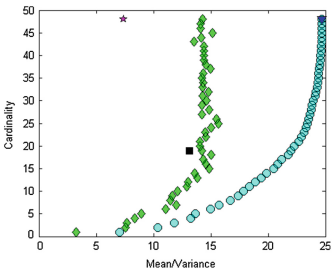


Fig. 3. Efficient frontier of the biobjective cardinality/mean-variance problem for DTS3. See the caption of Fig. 1 for an explanation of the various symbols.

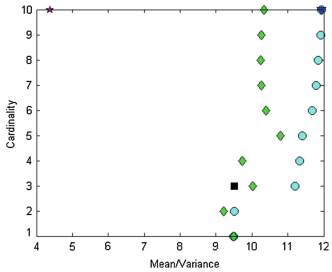


Fig. 4. Efficient frontier of the biobjective cardinality/mean-variance problem for FF10. ★ Naive ▼ Markowitz mean per variance ■ Markowitz minimum variance ♦ cardinality/mean-variance ♦ cardinality constrained minimum variance (Color figure online)

4.2 In-Sample Optimization

We then applied the solver `dms` (version 0.2) to compute the efficient frontier (or Pareto front) of the cardinality/mean-variance biobjective optimization problem (5). A few modifications to (5) were made before applying the solver as well as a few changes to the solver default parameters (the details are described in Appendix A). We present results for the initial in-sample optimization. For the FTSE 100 data sets this sample is from 01/2003 to 12/2006 and for the FF data sets is from 07/1971 to 06/1996. Figures 1, 2, 3, 4, 5, and 6 contain the plots of the efficient frontiers calculated for, respectively, the FTSE 100 and FF data sets. In all these plots we also marked three other portfolios. The first one is

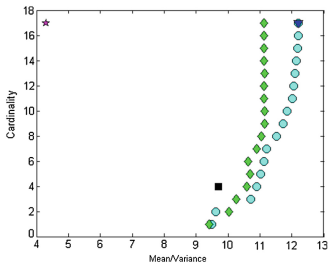


Fig. 5. Efficient frontier of the biobjective cardinality/mean-variance problem for FF17. See the caption of Fig. 4 for an explanation of the various symbols.

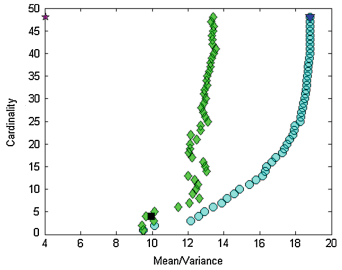


Fig. 6. Efficient frontier of the biobjective cardinality/mean-variance problem for FF48. See the caption of Fig. 4 for an explanation of the various symbols.

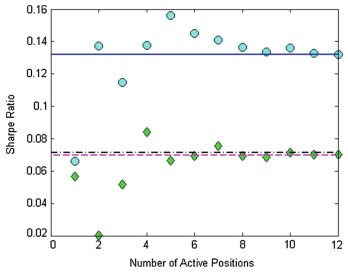


Fig. 7. Out-of-sample performance for DTS1 measured by the Sharpe ratio over all the out-of-sample periods. - - Naive — Markowitz mean per variance **-.-** Markowitz minimum variance • cardinality/mean-variance ♦ cardinality constrained minimum variance (Color figure online)

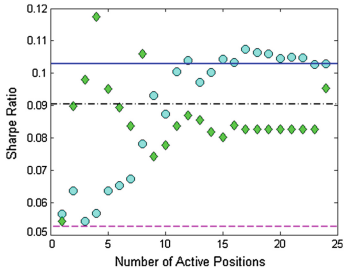


Fig. 8. Out-of-sample performance for DTS2 measured by the Sharpe ratio over all the out-of-sample periods. See the caption of Fig. 7 for an explanation of the various symbols and lines.

the $1/N$ portfolio corresponding to the naive strategy. A second one is obtained maximizing expected return per variance.

$$\begin{array}{ll} \min\limits_{w\in\mathbb{R}^N} & -\frac{\mu^\top w}{w^\top Q w} \\ \text{subject to} & e^\top w = 1. \end{array} \tag{6}$$

This portfolio corresponds to the extreme point (of maximum cardinality) of the efficient frontier (or Pareto front) of the cardinality/mean-variance biobjective optimization problem (5). The third one is a classical Markowitz related portfolio and is obtained by minimizing variance under no short-selling

$$\begin{array}{ll} \min\limits_{w\in\mathbb{R}^N} & w^\top Q w \\ \text{subject to} & e^\top w = 1, \\ & w \geq 0. \end{array} \tag{7}$$

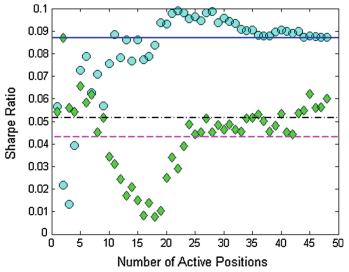


Fig. 9. Out-of-sample performance for DTS3 measured by the Sharpe ratio over all the out-of-sample periods. See the caption of Fig. 7 for an explanation of the various symbols and lines.

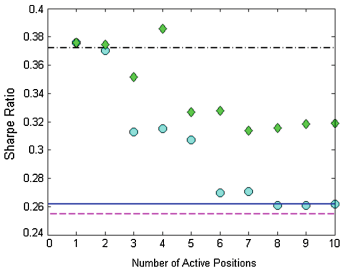


Fig. 10. Out-of-sample performance for FF10 measured by the Sharpe ratio over all the out-of-sample periods. - - Naive — Markowitz mean per variance **-.-** Markowitz minimum variance • cardinality/mean-variance ♦ cardinality constrained minimum variance (Color figure online)

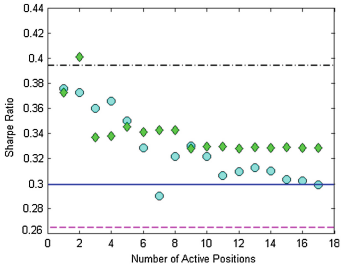


Fig. 11. Out-of-sample performance for FF17 measured by the Sharpe ratio over all the out-of-sample periods. See the caption of Fig. 10 for an explanation of the various symbols and lines.

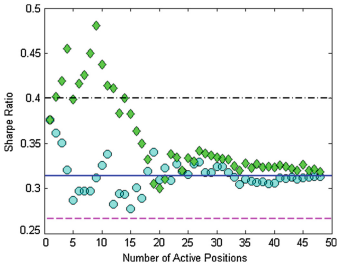


Fig. 12. Out-of-sample performance for FF48 measured by the Sharpe ratio over all the out-of-sample periods. See the caption of Fig. 10 for an explanation of the various symbols and lines.

This instance was solved using the `quadprog` function from the MATLAB [24] Optimization Toolbox. Regarding problem (7), it is known that not allowing short-sale has a regularizing effect on minimum-variance Markowitz portfolio selection (see [16]) and leads to portfolios of low cardinality.

Since we know that minimum variance portfolios outperform mean-variance portfolios (the estimate error of the expected returns is eliminated, see [16]), we considered the following cardinality constrained minimum variance model (instead of the one introduced in Sect. 2.3)

$$\begin{aligned} \min_{w \in \mathbb{R}^N} \quad & w^\top Q w \\ \text{subject to } & \text{card}(w) \leq K, \\ & e^\top w = 1, \\ & L_i \leq w_i \leq U_i, \quad i = 1, \ldots, N. \end{aligned}$$

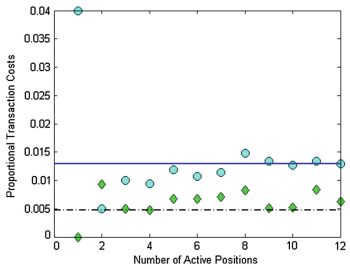


Fig. 13. Transaction costs of the efficient cardinality/mean-variance portfolios for DTS1. — Markowitz mean per variance **-.-** Markowitz minimum variance • cardinality/mean-variance ♦ cardinality constrained minimum variance (Color figure online)

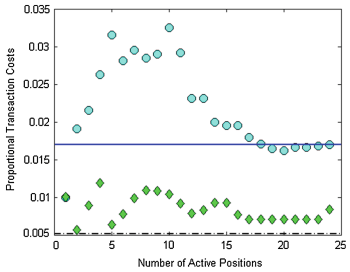


Fig. 14. Transaction costs of the efficient cardinality/mean-variance portfolios for DTS2. See the caption of Fig. 13 for an explanation of the various symbols and lines.

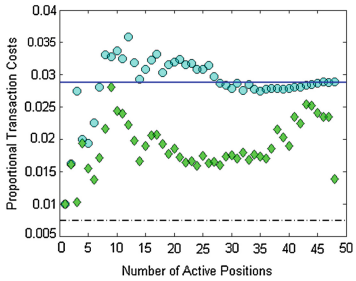


Fig. 15. Transaction costs of the efficient cardinality/mean-variance portfolios for DTS3. See the caption of Fig. 13 for an explanation of the various symbols and lines.

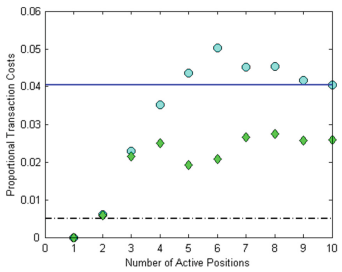


Fig. 16. Transaction costs of the efficient cardinality/mean-variance portfolios for FF10. — Markowitz mean per variance **-.-** Markowitz minimum variance • cardinality/mean-variance ♦ cardinality constrained minimum variance (Color figure online)

By introducing binary variables, one can rewrite this problem as a mixed-integer quadratic programming (MIQP) problem:

$$\begin{aligned} \min_{w,y\in\mathbb{R}^N} \quad & w^\top Q w \\ \text{subject to } & e^\top y \le K, \\ & e^\top w = 1, \\ & L_i y_i \le w_i \le U_i y_i, \quad i = 1,\dots,N, \\ & y_i \in \{0,1\}, \quad i = 1,\dots,N. \end{aligned} \tag{8}$$

We also mark in the plots the portfolios that result from solving problem (8) for each value of $K \in [1, N]$. For this purpose we used the solver `cplexmiqp` from ILOG IBM CPLEX for MATLAB [22].

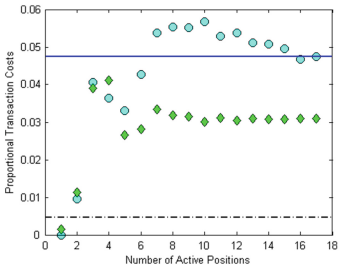


Fig. 17. Transaction costs of the efficient cardinality/mean-variance portfolios for FF17. See the caption of Fig. 16 for an explanation of the various symbols and lines.

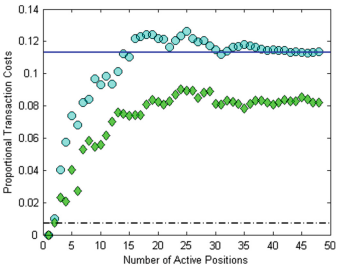


Fig. 18. Transaction costs of the efficient cardinality/mean-variance portfolios for FF48. See the caption of Fig. 16 for an explanation of the various symbols and lines.

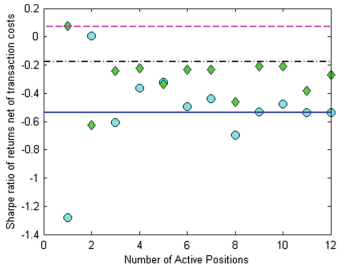


Fig. 19. Out-of-sample performance for DTS1 measured by the Sharpe ratio over all the out-of-sample periods. - - Naive — Markowitz mean per variance **-.-** Markowitz minimum variance • cardinality/mean-variance ♦ cardinality constrained minimum variance (Color figure online)

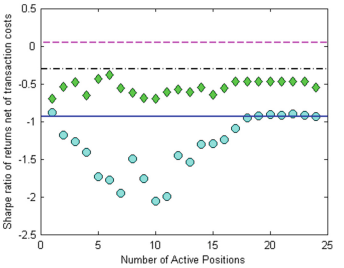


Fig. 20. Out-of-sample performance for DTS2 measured by the Sharpe ratio of returns net of transaction costs over all the out-of-sample periods.See the caption of Fig. 19 for an explanation of the various symbols and lines.

4.3 Out-of-sample Performance

The analysis of out-of-sample performance relies on a rolling-sample approach. For the FTSE 100 data sets we considered 12 periods (months) of evaluation. We begin by computing the efficient frontier (or Pareto front) of the cardinality/mean-variance biobjective optimization problem (5) for the in-sample time window from 01/2003 to 12/2006 (see Subsect. 4.2). We then held fixed each portfolio and observed its returns over the next period (January 2007). Then we discarded January 2003 and brought January 2007 into the sample. We repeated this process until exhausting the 12 months of 2007. We applied the same rolling-sample approach to the FF data sets, considering an initial in-sample time window from 07/1971 to 06/1996 (see Subsect. 4.2) and 15 periods of evaluation (the 15 next years).

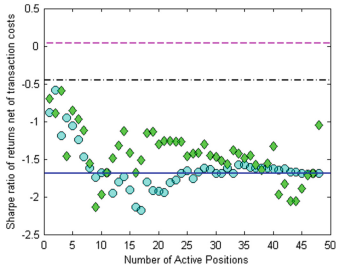


Fig. 21. Out-of-sample performance for DTS3 measured by the Sharpe ratio of returns net of transaction costs over all the out-of-sample periods.See the caption of Fig. 19 for an explanation of the various symbols and lines.

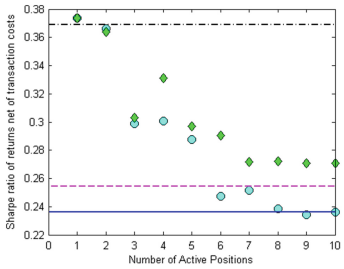


Fig. 22. Out-of-sample performance for FF10 measured by the Sharpe ratio of returns net of transaction costs over all the out-of-sample periods. - - Naive — Markowitz mean per variance **-.-** Markowitz minimum variance • cardinality/mean-variance ♦ cardinality constrained minimum variance (Color figure online)

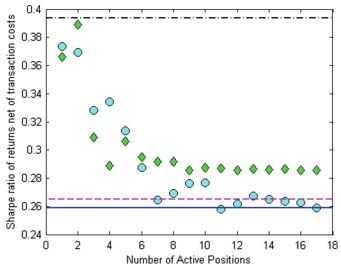


Fig. 23. Out-of-sample performance for FF17 measured by the Sharpe ratio of returns net of transaction costs over all the out-of-sample periods. See the caption of Fig. 22 for an explanation of the various symbols and lines.

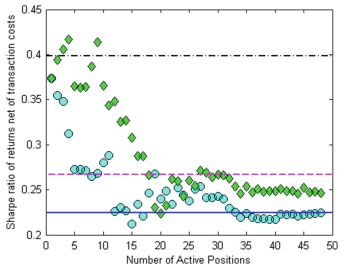


Fig. 24. Out-of-sample performance for FF48 measured by the Sharpe ratio of returns net of transaction costs over all the out-of-sample periods. See the caption of Fig. 22 for an explanation of the various symbols and lines.

4.4 Out-of-sample Performance Measured by the Sharpe Ratio

In each period of evaluation, the out-of-sample performance was then measured by the Sharpe ratio

$$S = \frac{m - r_f}{\sigma},$$

where m is the mean return, r_f is the return of the risk-free asset[1], and σ is the standard deviation. The results (over all the periods of evaluation) are given

[1] For the FTSE 100 data sets we used the 3 month Treasury-Bills UK. Such data is public and made available by the Bank of England, at the site http://www.bankofengland.co.uk. For the FF data sets we used the 90-day Treasury-Bills US. Such data is public and made available by the Federal Reserve, at the site http://www.federalreserve.gov.

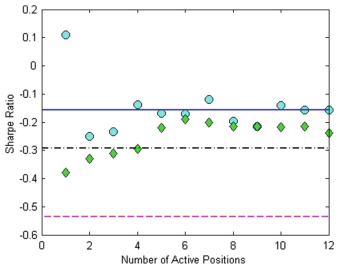


Fig. 25. Out-of-sample performance for DTS1, including the financial crisis years 2008–2010, measured by the Sharpe ratio over all the out-of-sample periods. - - Naive — Markowitz mean per variance **-.-** Markowitz minimum variance • cardinality/mean-variance ♦ cardinality constrained minimum variance (Color figure online)

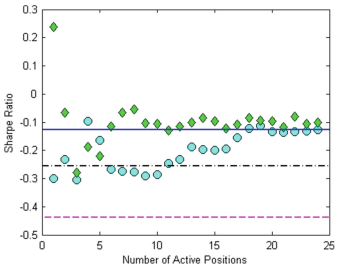


Fig. 26. Out-of-sample performance for DTS2, including the financial crisis years 2008–2010, measured by the Sharpe ratio over all the out-of-sample periods. See the caption of Fig. 25 for an explanation of the various symbols and lines.

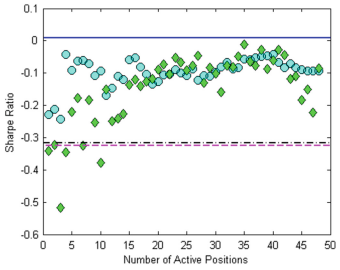


Fig. 27. Out-of-sample performance for DTS3, including the financial crisis years 2008–2010, measured by the Sharpe ratio over all the out-of-sample periods. See the caption of Fig. 25 for an explanation of the various symbols and lines.

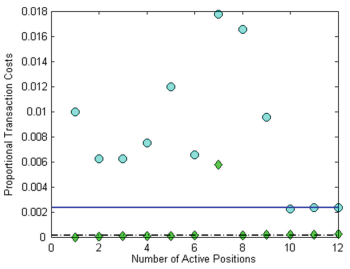


Fig. 28. Transaction costs of the efficient cardinality/mean-variance portfolios for DTS1, including the financial crisis years 2008–2010. — Markowitz mean per variance **-.-** Markowitz minimum variance • cardinality/mean-variance ♦ cardinality constrained minimum variance (Color figure online)

in Figs. 7, 8 and 9 for the FTSE 100 portfolios and in Figs. 10, 11 and 12 for the FF ones. Using IBM SPSS Statistics [23] we calculated the p-values for the statistical significance of the difference between Sharpe ratios of the benchmark naive portfolio and all the others computed portfolios. We did not report them here because they are not statistically significant.

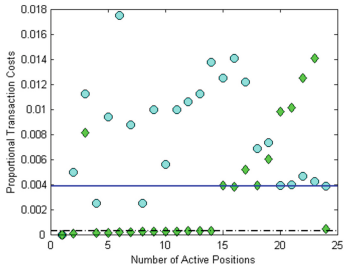


Fig. 29. Transaction costs of the efficient cardinality/mean-variance portfolios for DTS2, including the financial crisis years 2008–2010. See the caption of Fig. 28 for an explanation of the various symbols and lines.

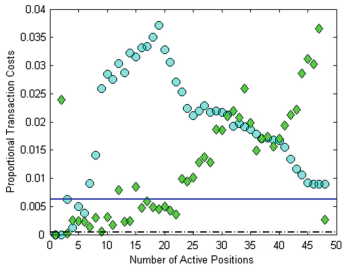


Fig. 30. Transaction costs of the efficient cardinality/mean-variance portfolios for DTS3, including the financial crisis years 2008–2010. See the caption of Fig. 28 for an explanation of the various symbols and lines.

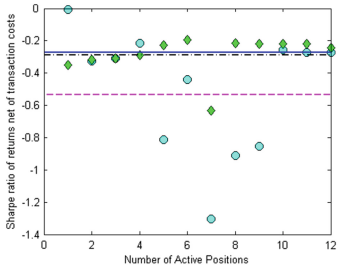


Fig. 31. Out-of-sample performance for DTS1, including the financial crisis years 2008–2010, measured by the Sharpe ratio of returns net of transaction costs over all the out-of-sample periods. - - Naive — Markowitz mean per variance **-.-** Markowitz minimum variance • cardinality/mean-variance ♦ cardinality constrained minimum variance (Color figure online)

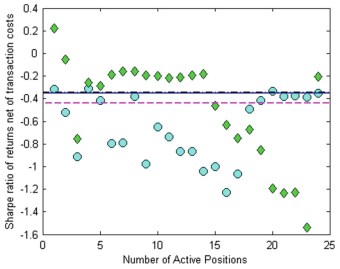


Fig. 32. Out-of-sample performance for DTS2, including the financial crisis years 2008–2010, measured by the Sharpe ratio of returns net of transaction costs over all the out-of-sample periods. See the caption of Fig. 31 for an explanation of the various symbols and lines.

4.5 Transaction Costs

Since one is rebalancing portfolios for each out-of-sample period, one can compute the transaction costs of such a trade. We set the proportional transaction cost equal to 50 basis points per transaction (as usually assumed in the literature). Thus the cost of a trade over all assets is given by

$$TC = \sum_{t=1}^{T-1} 0.5\,\% \sum_{i=1}^{N} \mid w_{i,t+1} - w_{i,t} \mid, \tag{9}$$

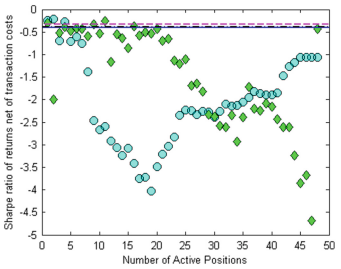


Fig. 33. Out-of-sample performance for DTS3, including the financial crisis years 2008–2010, measured by the Sharpe ratio of returns net of transaction costs over all the out-of-sample periods. See the caption of Fig. 31 for an explanation of the various symbols and lines.

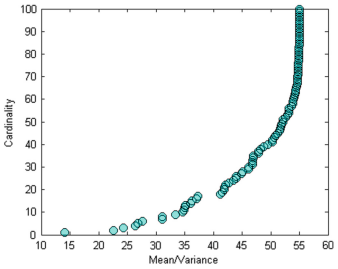


Fig. 34. Efficient frontier of the biobjective cardinality/mean-variance problem for FF100.

with $T = 12$ for the FTSE 100 data sets and $T = 15$ for the FF data sets. The results are given in Figs. 13, 14 and 15 for the FTSE 100 portfolios and in Figs. 16, 17 and 18 for the FF ones.

4.6 Out-of-sample Performance Measured by the Sharpe Ratio of Returns Net of Transaction Costs

In the presence of transaction costs we calculated the Sharpe ratio of returns net of transaction costs

$$SR = \frac{m - TC - r_f}{\sigma},$$

where m is the mean return, TC is the proportional transaction cost in (9), r_f is the return of the risk-free asset, and σ is the standard deviation. The out-of-sample performance was then measured by the Sharpe ratio of returns net of transaction costs. The results are given in Figs. 19, 20 and 21 for the FTSE 100 portfolios and in Figs. 22, 23 and 24 for the FF ones.

4.7 Results Including the Financial Crisis Years 2008–2010

The FTSE 100 data set used covered the period 2003–2007. With the aim of testing the robustness of the results, we also tried a FTSE 100 data set that covers the time window 2003–2010 (including thus the financial crisis years 2008–2010). The data sets were formed as described in Sect. 4.1, but excluding British Airways (see Table 1) due to missing data during the period considered, and including Wolseley (following an arbitrary alphabetic order). We performed an out-of-sample analysis as described in Sect. 4.3. We used daily periods of evaluation. We began by computing the efficient frontier (or Pareto front) of

the cardinality/mean-variance biobjective optimization problem (5) for the in-sample time window from 01/2003 to 12/2010 (using daily data). We then held fixed each portfolio and observed its returns over the next period (first trading day of January 2011). Then we discarded this first trading day of January 2011 and brought this into the sample. We repeated this process until exhausting the firsts 15 trading days of 2011.

The results of the out-of-sample performance measured by the Sharpe ratio[2] are given in Figs. 25, 26 and 27. The results of the proportional transaction costs, are given in Figs. 28, 29 and 30. The results of the out-of-sample performance measured by the Sharpe ratio of returns net of transaction costs, are given in Figs. 31, 32 and 33.

4.8 Discussion of the Results

Contrary to one could think, given the intractability of $f_2(w) = \text{card}(w)$ and the fact that no derivatives are being used for $f_1(w) = -\mu^\top w/w^\top Qw$, direct multisearch (the solver `dms`) was capable of quickly determining (in-sample) the efficient frontier for the biobjective optimization problem (5). For instance, for the data sets of roughly 50 assets, a regular laptop takes a few dozens of seconds to produce the efficient frontiers. We have a direct way of dealing with sparsity, which offers a complete determination of an efficient frontier for all cardinalities. According to *a priori* preferences, one could choose (in-sample) the desired cardinality. For the portfolios constructed using the FTSE 100 index data (portfolios of individual securities), a large number of our sparse portfolios, among the efficient cardinality/mean-variance ones, consistently overcame the naive strategy and at least one of the two related classical Markowitz models, in terms of out-of-sample performance measured by the Sharpe ratio. This effect has even happened for the largest data set (DTS3 with 48 securities), where the demand for sparsity is more relevant. For the portfolios constructed using the Fama/French benchmark collection (where securities are portfolios rather than individual securities), the scenario is different since the behavior of the naive strategy is even more difficult to outperform. Still, a large number of sparse efficient cardinality/mean-variance portfolios consistently overcame the naive strategy.

In both cases, FTSE 100 and FF data, the transaction costs of the efficient cardinality/mean-variance portfolios are lower than the mean per variance portfolio (solution of problem (6)) and higher than the minimum-variance portfolio (solution of problem (7)). Note that the minimum-variance portfolio does not allow short-selling, and so the weights at the outset are much more limited, thus leading to better results. Evaluating the performance out-of-sample by the Sharpe ratio of returns net of transaction costs (take into account the transaction costs), the efficient cardinality/mean-variance portfolios do not overcame the naive strategy for FTSE 100 data, but for FF data a large number of

[2] We used as a risk-free asset the daily startling certificate of deposit interest rate. Such data is public and made available by the Bank of England, at the site http://www.bankofengland.co.uk.

sparse efficient cardinality/mean-variance portfolios still consistently overcame the naive strategy. When we compare the performance results between the efficient cardinality/mean-variance portfolios and the cardinality constrained minimum variance portfolios (solution of (8)), without considering the transaction costs, we observed better results for the FTSE 100 and worse for the FF. The MIQP performed better in terms of Sharpe ratio of returns net of transaction costs since the cost of transaction costs are lower, one possible explanation for this is the fact of not taking into account the estimation of the expected returns. Moreover, our cardinality/mean-variance portfolios are truly efficient whereas the cardinality constrained minimum variance do not necessarily exhibit Pareto efficiency. For the FTSE 100 data set, the analysis including the financial crisis years 2008–2010 shows that the results are robust. Finally, we also computed the cardinality/mean-variance efficient frontier for the data set FF100, where portfolios are formed on size and book-to-market (see Fig. 34). (This time we needed a budget of the order of 10^7 function evaluations, see Appendix A.) We remark that FF48 and FF100 are the data sets also used in [8]. In this paper, as we said before, the authors focus on a modification to the Markowitz classical model by the incorporation of a term involving a multiple of the ℓ_1 norm of the vector of portfolio positions. Despite the different sparse-oriented techniques and different strategies for evaluating out-of-sample performance, in both approaches (theirs and ours), sparse portfolios are found overcoming the naive strategy. In our approach one computes sparse portfolios satisfying an efficient or nondominant property and one does it directly and in single run, whereas in [8], there is a need to vary a tunable parameter and select the portfolios according to some criterion to be met (for example, sparsity). It is unclear what sort of efficient or nondominant property their portfolios satisfy. Moreover, we provide results for all cardinality values (from 1 to 48 in FF48 and from 1 to 100 in FF100), while in [8] the authors report results for cardinality values from 4 and 48 (FF48) and from 3 to 60 (FF100). We therefore claim to have a more direct way of dealing with sparsity, which offers a complete determination of an efficient frontier for all cardinalities.

5 Conclusions and Perspectives for Future Work

In this paper we have developed a new methodology to deal with the computation of mean-variance Markowitz portfolios with pre-specified cardinalities. Instead of imposing a bound on the maximum cardinality or including a penalization or regularization term into the objective function (in classical Markowitz mean-variance models), we took the more direct approach of explicitly considering the cardinality as a separate goal. This led us to a cardinality/mean-variance biobjective optimization problem (5) whose solution is given in the form of an efficient frontier or Pareto front, thus allowing the investor to tradeoff among these two goals when having transaction costs and portfolio management in mind. In addition, and surprisingly, a significant portion of the efficient cardinality/mean-variance portfolios (with cardinality values considerably lower than the number N of securities) have exhibited superior out-of-sample performance (under

reasonably low transaction costs that only increase moderately with cardinality). We solved the biobjective optimization problem (5) using a derivative-free solver running direct multisearch. Direct-search methods based on polling are known in general to be slow but extremely robust due their directional properties. Such a feature is crucial given the difficulty of the problem (one discontinuous objective function, the cardinality, and discontinuous Pareto fronts). We have observed the robustness of direct multisearch, in other words, its capability of successfully solving a vast majority of the instances (all in our case) even if at the expense of a large budget of function evaluations. Direct multisearch was applied off-the-shelf to determine the cardinality/mean-variance efficient frontier. The structure of problem (5), or of its practical counterpart (10), was essentially ignored. One can use the fact that the first objective function is smooth and of known derivatives to speed up the optimization and reduce even further the budget of function evaluations. Moreover, we also point out that it is trivial to run the poll step of direct multisearch in a parallel mode.

The use of derivative-free single or multiobjective optimization opens the research range of future work in sparse or dense portfolio selection. In fact, since derivative-free algorithms only rely on zero order information, they are applicable to any objective function of black-box type. One can thus use any measure to quantify the profit and risk of a portfolio. The classical Markowitz model assumes that the return of a portfolio is a linear combination of the returns of the individual securities. Also, it implicitly assumes a Gaussian distribution for the return, letting its variance be a natural measure of risk. However, it is known from the analysis of stylized facts that the distribution for the return of securities exhibits tails which are fatter than the Gaussian ones. Practitioners consider other measures of risk and profit better tailored to reality. Our approach to compute the cardinality/mean-variance efficient frontier is ready for application in such general scenarios.

A Using Direct Multisearch to Determine Efficient Cardinality/Mean-Variance Portfolios

A few modifications to problem (5) were required to make it solvable by a multiobjective derivative-free solver, in particular by a direct multisearch one. In practice the first modification to (5) consisted of approximating the true cardinality, by introducing a tolerance ϵ,

$$\begin{aligned} \min_{w\in\mathbb{R}^N} \quad & -\frac{\mu^\top w}{w^\top Q w} \\ \min_{w\in\mathbb{R}^N} \quad & \sum_{i=1}^N \mathbb{1}_{\{|w_i|>\epsilon\}} \\ \text{subject to } & e^\top w = 1, \\ & L_i \le w_i \le U_i, \quad i = 1,\ldots,N. \end{aligned}$$

chosen as $\epsilon = 10^{-8}$ ($\mathbb{1}$ represents the indicator function). Secondly, we selected symmetric bounds on the variables $L_i = -b$ and $U_i = b$,

$$\begin{array}{ll} \min\limits_{w\in\mathbb{R}^N} & -\frac{\mu^\top w}{w^\top Q w} \\ \min\limits_{w\in\mathbb{R}^N} & \sum_{i=1}^{N} \mathbb{1}_{\{|w_i|>\epsilon\}} \\ \text{subject to} & e^\top w \;=\; 1, \\ & -b \;\leq\; w_i \;\leq\; b, \quad i=1,\ldots,N, \end{array}$$

setting $b = 10$. Finally, we eliminated the constraint $e^\top w = 1$ since direct search methods do not cope well with equality constraints. The version fed to the `dms` solver was then

$$\begin{array}{ll} \min\limits_{w(1:N-1)\in\mathbb{R}^{N-1}} & -\frac{\mu^\top w}{w^\top Q w} \\ \min\limits_{w(1:N-1)\in\mathbb{R}^{N-1}} & \sum_{i=1}^{N-1} \mathbb{1}_{\{|w_i|>\epsilon\}} \\ \text{subject to} & -b \;\leq\; w_i \;\leq\; b, \quad i=1,\ldots,N-1, \\ & -b \;\leq\; 1-\sum_{i=1}^{N-1} w_i \;\leq\; b, \end{array} \tag{10}$$

where w_N in $-\mu^\top w / w^\top Q w$ was replaced by $1 - \sum_{i=1}^{N-1} w_i$.

We used a ll the default parameters of `dms` (version 0.2) with the following four exceptions. First, we needed to increase the maximum number of function evaluations allowed (from 20000 to 2000000 for $N(= n)$ up to 50) given the dimension of our portfolios, as well as to require more accuracy by reducing the step size tolerance from 10^{-3} to 10^{-7}. Then we turned off the use of the cache of previously evaluated points to make the runs faster (the default version of `dms` keeps such a list to avoid evaluating points too close to those already evaluated). Lastly, we realized that initializing the list of feasible nondominated points with a singleton led to better results than initializing it with a set of roughly N points as it happens by default. Thus, we set the option `list` of `dms` to zero, which, given the bounds on the variables, assigns the origin to the initial list.

References

1. Anagnostopoulos, K.P., Mamanis, G.: A portfolio optimization model with three objectives and discrete variables. Comput. Oper. Res. **37**, 1285–1297 (2010)
2. Anagnostopoulos, K.P., Mamanis, G.: The mean-variance cardinality constrained portfolio optimization problem: An experimental evaluation of five multiobjective evolutionary algorithms. Expert Syst. Appl. **38**, 14208–14217 (2011)
3. Bach, F., Ahipasaoglu, S.D., d'Aspremont, A.: Convex relaxations for subset selection (2010). ArXiv 1006.3601
4. Benartzi, S., Thaler, R.H.: Naive diversification strategies in defined contribution saving plans. Am. Econ. Rev. **91**, 79–98 (2001)
5. Bertsimas, D., Shioda, R.: Algorithm for cardinality-constrained quadratic optimization. Comput. Optim. Appl. **43**, 1–22 (2009)
6. Bienstock, D.: Computational study of a family of mixed-integer quadratic programming problems. Math. Program. **74**, 121–140 (1996)
7. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)

8. Brodie, J., Daubechies, I., De Mol, C., Giannone, D., Loris, I.: Sparse and stable Markowitz portfolios. Proc. Natl. Acad. Sci. USA **106**, 12267–12272 (2009)
9. Cesarone, F., Scozzari, A., Tardella, F.: Efficient algorithms for mean-variance portfolio optimization with hard real-world constraints. Giornale dell'Istituto Italiano degli Attuari **72**, 37–56 (2009)
10. Cesarone, F., Scozzari, A., Tardella, F.: A new method for mean-variance portfolio optimization with cardinality constraints. Ann. Oper. Res. **205**, 213–234 (2013)
11. Chang, T.J., Meade, N., Beasley, J.E., Sharaiha, Y.M.: Heuristics for cardinality constrained portfolio optimisation. Comput. Oper. Res. **27**, 1271–1302 (2000)
12. Cornnuejols, G., Tütüncü, R.: Optimizations Methods in Finance. Cambridge University Press, Cambridge (2007)
13. DeMiguel, V., Garlappi, L., Nogales, F.J., Uppal, R.: A generalized approach to portfolio optimization: Improving performance by constrained portfolio norms. Manage. Sci. **55**, 798–812 (2009)
14. DeMiguel, V., Garlappi, L., Uppal, R.: Optimal versus naive diversification: How inefficient is the $1/N$ portfolio strategy? Rev. Financ. Stud. **22**, 1915–1953 (2009)
15. Fieldsend, J.E., Matatko, J., Peng, M.: Cardinality constrained portfolio optimisation. In: Yang, Z.R., Yin, H., Everson, R.M. (eds.) IDEAL 2004. LNCS, vol. 3177, pp. 788–793. Springer, Heidelberg (2004)
16. Jagannathan, R., Ma, T.: Risk reduction in large portfolios: why imposing the wrong constraints hfelps. J. Finan. **58**, 1651–1684 (2003)
17. Lin, D., Wang, S., Yan, H.: A multiobjective genetic algorithm for portfolio selection. Working paper. Institute of Systems Science, Academy of Mathematics and Systems Science Chinese Academy of Sciences, Beijing, China (2001)
18. Di Lorenzo, D., Liuzzi, G., Rinaldi, F., Schoen, F., Sciandrome, M.: A concave optimization-based approach for sparse portfolio selection. Optim. Methods Softw. **27**, 983–1000 (2012)
19. Markowitz, H.M.: Portfolio selection. J. Finan. **7**, 77–91 (1952)
20. Markowitz, H.M.: Portfolio Selection: Efficient Diversification of Investments. In: Cowles Foundation Monograph No 16. Wiley, New York (1959)
21. Steinbach, M.C.: Markowitz revisited: mean-variance models in financial portfolio analysis. SIAM Rev. **43**, 31–85 (2001)
22. IBM™. IBM ILOG CPLEX®
23. IBM™. IBM SPSS Statistics®
24. The MathWorks™. MATLAB®
25. Vielma, J.P., Ahmed, S., Nemhauser, G.L.: A lifted linear programming branch-and-bound algorithm for mixed-integer conic quadratic programs. INFORMS J. Comput. **20**, 438–450 (2008)
26. Woodside-Oriakhi, M., Lucas, C., Beasley, J.E.: Heuristic algorithms for the cardinality constrained efficient frontier. Eur. J. Oper. Res. **213**, 538–550 (2011)

Two Semi-Lagrangian Fast Methods for Hamilton-Jacobi-Bellman Equations

Simone Cacace[1], Emiliano Cristiani[2(✉)], and Maurizio Falcone[1]

[1] Dipartimento di Matematica, Sapienza – Università di Roma, Rome, Italy
[2] Istituto per le Applicazioni del Calcolo, CNR, Rome, Italy
e.cristiani@iac.cnr.it

Abstract. In this paper we apply the Fast Iterative Method (FIM) for solving general Hamilton–Jacobi–Bellman (HJB) equations and we compare the results with an accelerated version of the Fast Sweeping Method (FSM). We find that FIM can be indeed used to solve HJB equations with no relevant modifications with respect to the original algorithm proposed for the eikonal equation, and that it overcomes FSM in many cases. Observing the evolution of the active list of nodes for FIM, we recover another numerical validation of the arguments recently discussed in [1] about the impossibility of creating local single-pass methods for HJB equations.

Keywords: Single-pass methods · Fast iterative method · Fast sweeping method · Fast marching method

1 Introduction

The study of Hamilton–Jacobi (HJ) equations arises in several contexts, including classical mechanics, front propagation, control problems and differential games. In particular, for optimal control problems, the value function can be characterized as the unique viscosity solution of a Hamilton–Jacobi–Bellman (HJB) equation. Unfortunately, solving numerically the HJB equation can be rather expensive from the computational point of view. This is the reason why in the last years an increasing number of efficient techniques have been proposed, see, e.g., [1] for a brief review.

Basically, these algorithms are divided in two main classes: *single-pass* and *iterative*. An algorithm is said to be *single-pass* if one can fix *a priori* a (small) number r which depends only on the equation and on the mesh structure (not on the number of mesh points) such that each mesh point is re-computed at most r times. Single-pass algorithms usually divide the numerical grid in, at least, three time-varying subsets: *Accepted* (ACC), *Considered* (CONS), and *Far* (FAR).

This research was supported by the following grants: AFOSR Grant FA9550-10-1-0029, ITN-Marie Curie Grant 264735-SADCO.

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 74–84, 2014.
DOI: 10.1007/978-3-662-45504-3_7

Nodes in ACC are definitively computed, nodes in CONS are computed but their values are not yet final, and nodes in FAR are not yet computed. We say that a single-pass algorithm is *local* if the computation at any mesh point involves only the values of first neighboring nodes, the region CONS is 1-cell-thick and no information coming from FAR region is used. The methods which are not single-pass are iterative.

Among fast methods, the prototype algorithm for the local single-pass class is the Fast Marching Method (FMM) [9,12], while that for the iterative class is the Fast Sweeping Method (FSM) [7,8,11,13]. Another interesting method is the Fast Iterative Method (FIM) [4–6], which shares some features with both iterative and single-pass methods. Recently, Cacace et al. [1] have shown that it is not possible to create a local single-pass algorithm for solving general HJB equations. This motivates the efforts to develop new techniques, particularly in the class of iterative methods.

In this paper, we focus on the following minimum time HJB equation

$$\sup_{a\in B_1} \{-f(x,a)\cdot \nabla T(x)\} = 1, \qquad x \in \mathbb{R}^d \backslash \mathcal{T} \tag{1}$$

where d is the space dimension, $\mathcal{T}$ is a closed nonempty target set in $\mathbb{R}^d$, $f : \mathbb{R}^d \times B_1 \to \mathbb{R}^d$ is a given vector-valued Lipschitz continuous function, and B_1 is the unit ball in $\mathbb{R}^d$ centered in the origin, representing the set of the admissible controls. We complement the equation with homogeneous Dirichlet condition $T = 0$ on $\mathcal{T}$. Let us note that if $f(x,a) = c(x)a$, Eq. (1) becomes the eikonal equation $c(x)|\nabla T(x)| = 1$. To simplify the notations, we restrict the discussion to the case $d = 2$. Generalizations of the considered algorithms to any space dimension is straightforward, although the implementation is not trivial.

The goal of this paper is twofold: First, we investigate the possibility of applying a semi-Lagrangian version of the FIM to Eq. (1). To our knowledge, FIM was only used for solving the eikonal equation [5] and a special class of HJ equations [6], although there is no particular constraint to apply it in a more general framework. The algorithm indeed does not rely on the special form and features of the eikonal equation. In addition, we measure the degree of "iterativeness" of FIM, keeping track of how many times each grid node is inserted into the list of nodes which are actually computed at each step. Interestingly, the results indirectly confirm the findings of [1], showing that general HJB equations require the nodes to be visited an (*a priori*) unknown number of times, i.e. single-pass methods do not apply.

Second, we propose a new acceleration technique for the FSM, which is effective when (1) is discretized by means of a semi-Lagrangian scheme (see [3] for a comprehensive introduction). It reduces the CPU load for the sup search in (1), neglecting the control directions which are downwind with respect to the current sweep. The new method results to be remarkably faster, although (in general) the number of iterations needed for convergence increases.

2 Semi-Lagrangian Approximation

Let us introduce a structured grid G and denote its nodes by x_i, $i = 1, \ldots, N$. The space step is assumed to be uniform and equal to $\Delta x > 0$. Standard arguments [3] lead to the following discrete version of Eq. (1):

$$T(x_i) \approx \widehat{T}(x_i) = \min_{a \in B_1} \left\{ \widehat{T}(\tilde{x}_{i,a}) + \frac{|x_i - \tilde{x}_{i,a}|}{|f(x_i, a)|} \right\}, \qquad x_i \in G \tag{2}$$

where $\tilde{x}_{i,a}$ is a *non-mesh* point, obtained by integrating, until a certain final time τ, the ordinary differential equation

$$\begin{cases} \dot{y}(t) = f(y, a), & t \in [0, \tau] \\ y(0) = x_i \end{cases} \tag{3}$$

and then setting $\tilde{x}_{i,a} = y(\tau)$. To make the scheme fully discrete, the set of admissible controls B_1 is discretized with N_c points and we denote by a^* the optimal control achieving the minimum. Note that we can get different versions of the semi-Lagrangian (SL) scheme (2) varying τ, the method used to solve (3), and the interpolation method used to compute $\widehat{T}(\tilde{x}_{i,a})$. Moreover, we remark that, in any single-pass method, the computation of $\widehat{T}(x_i)$ cannot involve the value $\widehat{T}(x_i)$ itself, because this self-dependency would make the method iterative. Here we use a 3-point scheme: Eq. (3) is solved by an explicit forward Euler scheme until the solution is at distance Δx from x_i, where it falls inside the triangle of vertices $x_{i,1}$, $x_{i,2}$, and $x_{i,3}$, to be chosen among the first neighbors of x_i. The value $\widehat{T}(x_i)$ is computed by a two-dimensional linear interpolation of the values $\widehat{T}(x_{i,1})$, $\widehat{T}(x_{i,2})$ and $\widehat{T}(x_{i,3})$ (see [1] for details).

3 Limits of Local Single-Pass Methods

In this section we briefly recall the main result of [1]. From the numerical point of view, it is meaningful to divide HJB equations into four classes. For any given mesh, we have:

(ISO) Equations whose characteristic curves coincide or lie in the same simplex of the gradient curves of their solutions.

(¬ISO) Equations for which there exists at least a grid node where the characteristic curve and the gradient curve of the solution do not lie in the same simplex.

(REG) Equations with non-crossing (regular) characteristic curves. Characteristics spread from the target $\mathcal{T}$ to the rest of the domain without intersecting.

(¬REG) Equations with crossing characteristic curves. Characteristics start from the target $\mathcal{T}$ and then meet in finite time, creating shocks. As a result, the solution T is not differentiable at shocks.

Let us summarize here the main remarks on single-pass methods:
• FMM works for equations of type ISO and fails for equations of type ¬ISO (see [10] for further details and explanations), while FSM can be successfully applied in any case.
• Handling ¬ISO case requires CONS not to follow the level sets of the solution itself. Indeed, if CONS turns out to be an approximation of the level sets of the solution, it means that the solution is computed in an increasing order, thus following the gradient curve rather than the characteristic curve.
• Handling ¬REG case requires CONS to be an approximation of the level sets of the solution. Let us clarify this point. Consider the ¬REG case and let x be a point belonging to a shock, i.e. where the solution is not differentiable. By definition, the value $T(x)$ is carried by two or more characteristic curves reaching x at the same time. Similarly, let x_i be a grid node Δx-close to the shock. In order to mimic the continuous case, x_i has to be approached by the ACC region approximately at the same time from the directions corresponding to the characteristic curves. In this case, the value $T(x_i)$ is correct (no matter which upwind direction is chosen) and, more important, the characteristic information stops at x_i and it is no longer propagated, getting stuck by the ACC region. As a consequence, the shock is localized properly. On the other hand, if CONS region is not an approximation of a level set of the solution a node x_i close to a shock can be reached by ACC at different times. When ACC reaches x_i for the first time, it is impossible to detect the presence of the shock by using only local information. Indeed, only a global view of the solution allows one to know that another characteristic curve will reach x_i at a later time. As a consequence, the algorithm continues the enlargement of CONS and ACC, thus making an error that cannot be corrected by the following iterations.

In conclusion, we get that local single-pass methods cannot handle equations ¬ISO & ¬REG. In this situation, one has to add non local information regarding the location of the shock, or going back to nodes in ACC at later time, breaking the single-pass property. This motivates the investigation of new techniques, especially iterative methods, as the ones described in the next sections.

4 Fast Iterative Method

In this section we briefly recall the construction of FIM [4–6]. As in FMM, the main idea of FIM is to update only few grid nodes at each step. These nodes are stored in a separated list, called *active* list. During each step, the list of active nodes is modified, and the band thickens or expands to include all nodes that could be affected by the current updates. A node can be removed from the active list when its value is up to date with respect to its neighbors (i.e., it has reached convergence) and can be appended to the list (relisted) whenever any upwind neighbor's value has changed.

FIM is formally an iterative method, since the number of times a grid node is visited depends on the dynamics and on the grid size. On the other hand, the active list resembles the set CONS of FMM, and in some special cases FIM is in

fact a single-pass algorithm, see Sect. 6. Nevertheless, the active list and CONS differ for some important features. The first is that the active list is not kept ordered, and then the causality relationship among grid nodes is lost. The second is that the active list can be more than 1-node-thick, i.e. it can approximate a two-dimensional set. Finally, grid nodes removed from the active list can re-enter at a later time. This is the price to pay for loosing the causality.

The FIM algorithm consists of two parts, the initialization and the updating. In the initialization step, one has to set the boundary conditions and set the values of the rest of the grid nodes to infinity (or some very large value). Next, the adjacent neighbors of the source nodes (i.e. the target) are added to the active list. In the updating step, for every point in the list, one computes the new value and checks if the value at the node has converged by comparing the old and the new value at the considered point. If it has converged, one removes the node from the list and append to the list any non active adjacent node such that its updated value is less than the current one. The algorithm runs until the list is empty.

FIM was introduced for solving a special class of HJ equations [5,6]. Nevertheless, in Sect. 6 we show that FIM based on a SL discretization *can be successfully applied to general HJB equations with no modifications.*

5 An Optimized Fast Sweeping Method

FSM is another popular method for solving HJ equations [7,8,11,13]. The main advantage of the method is its implementation, which is extremely easy (easier than that of FMM and FIM). FSM is basically the classical iterative (fixed-point) method, since each node is visited in a predefined order, until convergence is reached. Here, the visiting directions (sweeps) are alternated in order to follow all possible characteristic directions, trying to exploit causality. In two-dimensional problems, the grid is visited sweeping in four directions: $S \rightarrow N \,\&\, W \rightarrow E$, $S \rightarrow N \,\&\, E \rightarrow W$, $N \rightarrow S \,\&\, E \rightarrow W$ and $N \rightarrow S \,\&\, W \rightarrow E$.
The key point is the Gauss-Seidel-like update of grid nodes, which allows one to compute in a cascade fashion a relevant part of the grid nodes in only one sweep. Indeed it is well known that in the case of eikonal equations FSM converges in only four sweeps [13].

Here we propose an easy modification of the FSM based on a SL discretization, aiming at saving CPU time for each sweep. Let us explain the idea in the case of a dynamics of the form $c(x,a)a$, with $c > 0$. It is clear that during the sweep $S \rightarrow N \,\&\, W \rightarrow E$ the algorithm cannot exploit the power of the Gauss-Seidel cascade for the information coming from NE. Indeed, even if a node actually depends on its NE neighbor, that information flows upwind and it is not propagated to other nodes during the current sweep. Then, we propose to *remove downwind discrete controls from the minimum search in the SL scheme* (2), since they have small or no effect in the update of the nodes, see Fig. 1.
The assumption $c > 0$ is needed to preserve the order of the quadrants between the control a and the resulting dynamics $c(x,a)a$. Otherwise, the choice of controls to be removed should be adapted according to the sign of c.

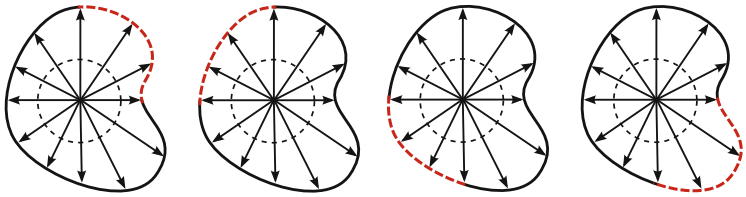


Fig. 1. Downwind controls with respect to the four sweeps: dashed arcs identify the directions to be removed

Note that the control set B_1 is reduced to a upwind 3/4 of ball. Then, let us denote by Upwind Fast Sweeping Method 3/4 (UFSM3/4) the classical FSM with this control reduction. An additional speedup, that we expect to work only in case where the characteristics are essentially straight, consists in reducing further the control ball to a upwind 1/4 of ball (UFSM1/4).

6 Numerical Experiments

In this section we compare the performance of FSM, FIM, UFSM1/4 and UFSM3/4 on the following equations:

Equation	Dynamics	Class
HJB-1	$f(x,y,a) = a$	ISO & REG
HJB-2	$f(x,y,a) = (1 + 4\chi_{\{x>1\}})\, a$	ISO & ¬REG
HJB-3	$f(x,y,a) = m_{\lambda,\mu}(a)\, a$	¬ISO & REG
HJB-4	$f(x,y,a) = F_2(x,y) m_{p(x,y),q(x,y)}(a)\, a$	¬ISO & ¬REG
HJB-5	$f(x,y,a) = (1 + \lvert x+y\rvert) m_{\lambda,\mu}(a)\, a$	¬ISO & ¬REG

where we defined $m_{\lambda,\mu}(a) = (1 + (\lambda\, a_1 + \mu\, a_2)^2)^{-\frac{1}{2}}$ for $\lambda, \mu \in \mathbb{R}$ and we denoted by χ_S the characteristic function of a set S. Moreover, for $c_1, c_2, c_3, c_4 > 0$, we defined

$$C(x) = c_1 \sin\left(\frac{c_2 \pi x}{c_3} + c_4\right), \qquad \big(F_1(x,y), F_2(x,y)\big) = \begin{cases} (0.5, 1) & \text{if } y \le C(x) \\ (2, 3) & \text{otherwise} \end{cases},$$

$$M(x,y) = \sqrt{\frac{\frac{F_2^2(x,y)}{F_1^2(x,y)} - 1}{1 + C'^2(x)}}, \quad p(x,y) = M(x,y)C'(x), \quad q(x,y) = -M(x,y)\,.$$

In all the following tests we set $\Omega = [-2,2]^2$ (except Test 4), the target $\mathcal{T} = (0,0)$ and the number of discrete controls $N_c = 32$. Regarding FIM, we keep track of the history of the active list by counting the number I_i of times the node x_i enters the active list. The number $I_{\max} := \max_i I_i$ gives a measure of the "iterativeness" of the method.

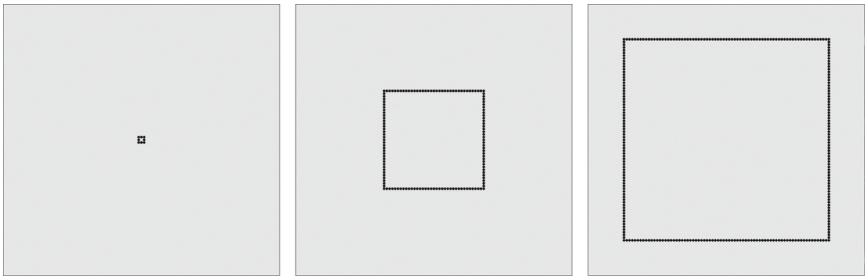


Fig. 2. FIM for HJB-1: active nodes at different steps

Test 1. Here we solve equation HJB-1. Figure 2 shows the evolution of FIM's active list. Differently from FMM, where the CONS set expands from the target following concentric circles (i.e. the level sets of the solution), here the active set moves following concentric squares (cf. the behavior of the CONS region of the *safe method* studied in [1]). As one can expect $I_{\max} = 1$, meaning that FIM behaves like a single-pass method. Table 1 compares CPU times of the methods on different grids and the number of sweeps needed by sweeping methods to reach convergence. FSM converges in 4 sweeps for this equation, the additional sweep reported in Table 1 is the one required by the algorithm to check convergence. All the methods compute the same solution. In particular we see that FIM is slightly slower than FSM, as noted in [5]. On the other hand, UFSM methods (both 1/4 and 3/4) still converge in 5 sweeps, thus overcoming FSM.

Test 2. Here we solve equation HJB-2. Figure 3 shows the optimal vector field $f(x, a^*)$ and the history of active nodes (in grey scale, where black corresponds to $I_{\max}$ and white to 0). The maximal number of re-activation is $I_{\max} = 3$ and re-activation of nodes appears for the first time close to the shock line, see Fig. 3-center. This depends on the fact that the active list is not an approximation of

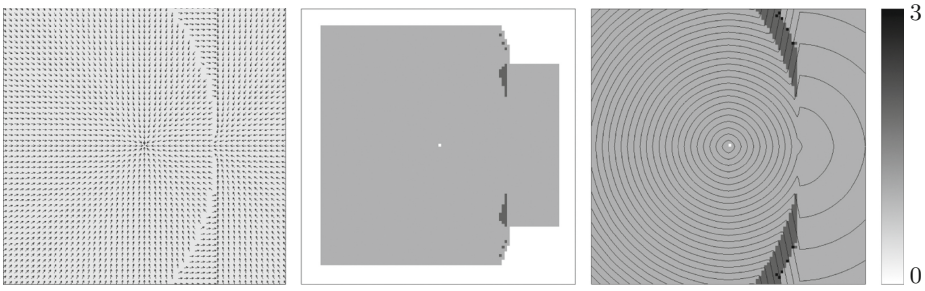


Fig. 3. FIM for HJB-2: Optimal controls (left), re-activation history at an intermediate step (center), re-activation history and level sets of the solution (right)

a level set of the solution and the equation HJB-2 falls in the ¬REG class. Then FIM is not able to capture the shock properly (see Sect. 3) in a single-pass fashion, but has to come back to recompute wrong values. Table 1 compares the methods. Results are similar to those of the previous test.

Test 3. Here we solve equation HJB-3 for $\lambda = 10$ and $\mu = 5$, namely the anisotropic eikonal equation, a well known example where FMM fails in computing the correct solution, due to the fact that characteristics do not coincide with gradient curves of the solution, see [10]. In this case FIM let evolve the active list as for the eikonal equation (Test 1, Fig. 2) and produces a maximal number of re-activation $I_{\max} = 1$. Again, this means that equation HJB-3 can be successfully solved by a local single-pass method, as the *safe method* introduced in [1] for the class REG. We refer to Table 1 for a comparison of the methods.

Test 4. Here we solve equation HJB-4 in $\Omega = [-0.5, 0.5]^2$ for $c_1 = 0.1225$, $c_2 = 2$, $c_3 = 0.5$ and $c_4 = 0$, an example of class ¬ISO & ¬REG coming from seismic imaging. It is a inhomogeneous anisotropic eikonal equation on a domain with two layers separated by a sinusoidal profile $C(x)$, with different constant anisotropy coefficients in each layer (given by the pairs $(F_1, F_2) = (0.5, 1)$ and $(F_1, F_2) = (2, 3)$).

All the methods compute the same solution, meaning that FIM can work for equations with substantial anisotropy and inhomogeneities (see also next test). Unexpectedly, also UFSM1/4 is able to correctly follow quite curved characteristics, see Fig. 4-left). Results in Table 1 show that sweeping methods need a large number of sweeps to reach convergence (even more for UFSMs, due to the control set reduction). This makes FIM be the fastest method. The maximal number of re-activation for the active list is $I_{\max} = 7$ and Fig. 4-center/right shows that re-activation of nodes appears both close to the shocks and where the optimal field exhibits rapid changes of direction.

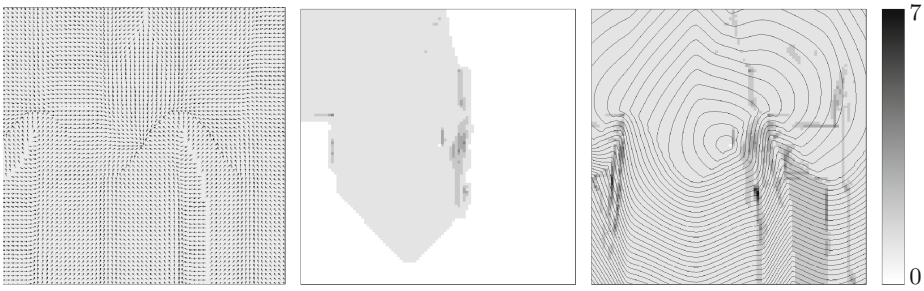


Fig. 4. FIM for HJB-4: Optimal controls (left), re-activation history at an intermediate step (center), re-activation history and level sets of the solution (right)

Test 5. Here we solve HJB-5 for $\lambda = 10$ and $\mu = 5$. This is the hardest example of class ¬ISO & ¬REG presented in [1], where the shock (see the cubic-like curve in Fig. 5-left/center) and a strong anisotropy region meet at the target. Sweeping methods FSM and UFSM3/4 require much more sweeps with respect to the previous tests, while UFSM1/4 fails in computing the correct solution, confirming that the control set reduction to 1/4 of ball cannot be applied in any case.

The maximal number of re-activation for FIM is $I_{\max} = 30$ (see Fig. 5-right), whereas the evolution of the active list is extremely complicated and also produces *regions of dimension two* (see Fig. 6). Nevertheless, results in Table 1 shows that, as the grid increases, FIM is still the fastest method. The presence of a two-dimensional active list clearly proves that local single-pass methods cannot be applied since the enlargement of CONS is required (cf. the buffered fast marching method [2]).

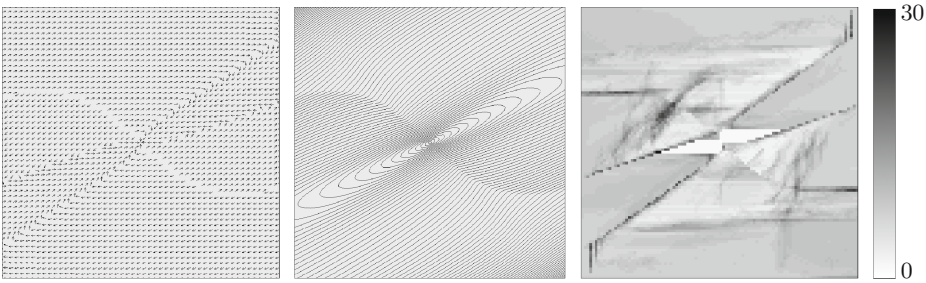


Fig. 5. FIM for HJB-5: Optimal controls (left), level sets of the solution (center), re-activation history (right)

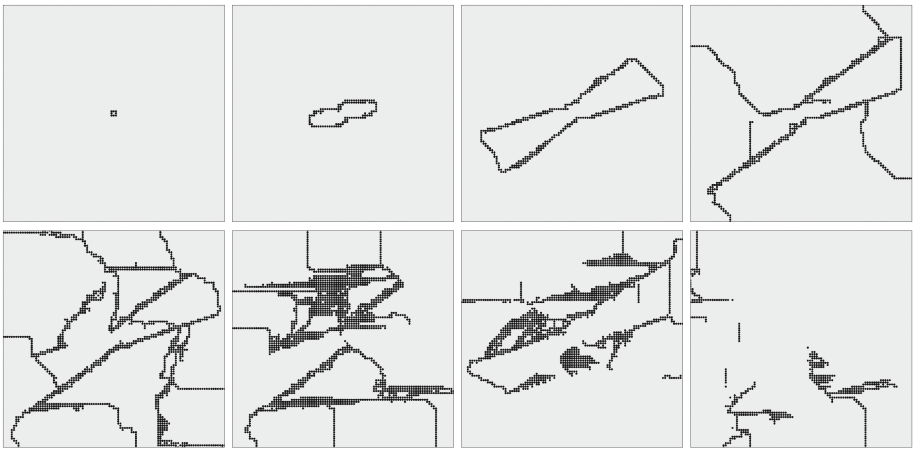


Fig. 6. FIM for HJB-5: active nodes at different steps

Table 1. CPU times (seconds) and number of sweeps. Fastest method in bold

Equation	Grid	Δx	FSM (sweeps)	FIM	UFSM1/4 (sweeps)	UFSM3/4 (sweeps)
HJB-1	101	0.04	0.13 (5)	0.17	**0.04** (5)	0.10 (5)
HJB-1	201	0.02	0.51 (5)	0.72	**0.15** (5)	0.40 (5)
HJB-1	401	0.01	2.04 (5)	3.21	**0.58** (5)	1.51 (5)
HJB-2	101	0.04	0.16 (5)	0.21	**0.04** (5)	0.12 (5)
HJB-2	201	0.02	0.63 (5)	0.87	**0.19** (5)	0.46 (5)
HJB-2	401	0.01	2.46 (5)	3.80	**0.70** (5)	1.84 (5)
HJB-3	101	0.04	0.31 (5)	0.38	**0.09** (5)	0.23 (5)
HJB-3	201	0.02	1.23 (5)	1.56	**0.35** (5)	0.88 (5)
HJB-3	401	0.01	4.88 (5)	6.57	**1.38** (5)	3.53 (5)
HJB-4	101	0.01	5.72 (25)	**1.93**	2.18 (34)	5.62 (34)
HJB-4	201	0.005	22.70 (25)	**7.68**	8.66 (34)	19.58 (30)
HJB-4	401	0.0025	99.38 (28)	**29.36**	34.07 (34)	77.14 (30)
HJB-5	101	0.04	3.23 (53)	5.30	-	**2.38** (53)
HJB-5	201	0.02	13.30 (55)	**3.32**	-	9.78 (55)
HJB-5	401	0.01	52.93 (55)	**14.53**	-	39.16 (55)

7 Conclusions

Tests performed in Sect. 6 show that FIM can be successfully used for solving general HJB equations with no modifications with respect to the original algorithm. Moreover, FIM appears to be the fastest method in case of complicated ¬ISO & ¬REG equations because of the large number of iterations needed by the sweeping methods. Considering that the implementation of FIM is not harder than that of FSM, we think that, overall, FIM is the best method among the tested ones.

UFSM3/4 is always preferable to FSM, since the larger number of iterations needed for convergence are widely counterbalanced by the speedup for each single sweep. UFSM1/4 is instead not safely applicable for general equations.

Finally, the results of this paper confirm those of [1]. Complicated ¬ISO & ¬REG equations require to pass through some nodes more than one time (cf., e.g., Fig. 5-right and [1, Fig. 7]), and exhibit two-dimensional regions in which every node depends on each other.

References

1. Cacace, S., Cristiani, E., Falcone, M.: Can local single-pass methods solve any stationary Hamilton-Jacobi-Bellman equation? SIAM J. Sci. Comput. **36**, A570–A587 (2014)

2. Cristiani, E.: A fast marching method for Hamilton-Jacobi equations modeling monotone front propagations. J. Sci. Comput. **39**, 189–205 (2009)
3. Falcone, M., Ferretti, R.: Semi-Lagrangian Approximation Schemes for Linear and Hamilton-Jacobi Equations. SIAM, Philadelphia (2014)
4. Fu, Z., Jeong, W.-K., Pan, Y., Kirby, R.M., Whitaker, R.T.: A fast iterative method for solving the eikonal equation on triangulated surfaces. SIAM J. Sci. Comput. **33**, 2468–2488 (2011)
5. Jeong, W.-K., Whitaker, R.T.: A fast iterative method for eikonal equations. SIAM J. Sci. Comput. **30**, 2512–2534 (2008)
6. Jeong, W.-K., Whitaker, R.T.: A fast iterative method for a class of Hamilton-Jacobi equations on parallel systems. University of Utah, Technical report UUCS-07-010 (2007)
7. Kao, C.Y., Osher, S., Qian, J.: Lax-Friedrichs sweeping scheme for static Hamilton-Jacobi equations. J. Comput. Phys. **196**, 367–391 (2004)
8. Qian, J., Zhang, Y.-T., Zhao, H.-K.: A fast sweeping method for static convex Hamilton-Jacobi equations. J. Sci. Comput. **31**, 237–271 (2007)
9. Sethian, J.A.: A fast marching level set method for monotonically advancing fronts. Proc. Natl. Acad. Sci. USA **93**, 1591–1595 (1996)
10. Sethian, J.A., Vladimirsky, A.: Ordered upwind methods for static Hamilton-Jacobi equations: theory and algorithms. SIAM J. Numer. Anal. **41**, 325–363 (2003)
11. Tsai, Y., Cheng, L., Osher, S., Zhao, H.: Fast sweeping algorithms for a class of Hamilton-Jacobi equations. SIAM J. Numer. Anal. **41**, 673–694 (2004)
12. Tsitsiklis, J.N.: Efficient algorithms for globally optimal trajectories. IEEE Trans. Autom. Control **40**, 1528–1538 (1995)
13. Zhao, H.: A fast sweeping method for eikonal equations. Math. Comput. **74**, 603–627 (2005)

Dynamic Sampling Schemes for Optimal Noise Learning Under Multiple Nonsmooth Constraints

Luca Calatroni[1(✉)], Juan Carlos De Los Reyes[2], and Carola-Bibiane Schönlieb[3]

[1] Cambridge Centre for Analysis, University of Cambridge, Cambridge, UK
lc524@cam.ac.uk
[2] Centro de Modelización Matemática, EPN Quito, Quito, Ecuador
[3] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK

Abstract. We consider the bilevel optimisation approach proposed in [5] for learning the optimal parameters in a Total Variation (TV) denoising model featuring for multiple noise distributions. In applications, the use of databases (dictionaries) allows an accurate estimation of the parameters, but reflects in high computational costs due to the size of the databases and to the nonsmooth nature of the PDE constraints. To overcome this computational barrier we propose an optimisation algorithm that, by sampling *dynamically* from the set of constraints and using a quasi-Newton method, solves the problem accurately and in an efficient way.

1 Introduction

Most images in the real world suffer from noise. In photography noisy images occur when taking a photograph under bad lighting conditions, for instance. Medical imaging applications, such as Magnetic Resonance Imaging (MRI) and Positron Electron Tomography (PET), produce under-sampled and noisy image data. In general, the quality of images obtained from imaging devices in the real world, in the sciences and medicine, is limited by the hardware and the limited time available to measure the image data. Hence, one of the most important tasks in image processing is the reduction of noise in images, called image denoising.

A common challenge in image denoising is the setup of a suitable denoising model. The model depends on the noise distribution and the class of images the denoised solution should belong to. In [5] a bilevel optimisation approach to learn the correct setup for a TV denoising model from a set of noisy and clean test images is proposed. There, optimal parameters $\lambda_i \in \mathbb{R}, i = 1, \ldots, d$ are determined by solving the following optimisation problem:

This work is supported by the King Abdullah University for Science and Technology (KAUST) Award No. KUK-I1-007-43, the EPSRC first grant Nr. EP/J009539/1 and the Cambridge Center for Analysis (CCA).

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 85–95, 2014.
DOI: 10.1007/978-3-662-45504-3_8

$$\min_{\lambda_i \geq 0,\ i=1,\dots,d} \frac{1}{2N} \sum_{k=1}^{N} \|\hat{u}_k - u_k\|^2_{L^2(\Omega)} \tag{1.1}$$

subject to the set of nonsmooth constraints:

$$\hat{u}_k = \text{argmin}_{u \in BV(\Omega) \cap \mathcal{A}} \left(|Du|(\Omega) + \sum_{i=1}^{d} \lambda_i \int_\Omega \phi_i(u, f_k)\, dx \right), \quad k = 1, \dots, N. \tag{1.2}$$

In (1.1)–(1.2) $\Omega \subset \mathbb{R}^2$ is the image domain, $|Du|(\Omega)$ is the Total Variation (TV) of u in Ω and $BV(\Omega)$ is the space of functions of bounded variation (see [1]). For each k, the pair (u_k, f_k) is an element of a set of N pairs of clean and noisy test images, respectively, whereas $\hat{u}_k$ is the TV-denoised version of f_k. For $i = 1, \dots, d$ the terms ϕ_i represent the different data fidelities, each one modelling one particular type of noise weighted by a parameter λ_i. Examples of ϕ are $\phi(u, f_k) = (u - f_k)^2$ for noise with Gaussian distribution and $\phi(u, f_k) = |u - f_k|$ for the case of impulse noise. The set $\mathcal{A}$ is the set of admissible functions such that the data fidelity terms are well defined.

In this paper, we use a simulated database of clean and noisy images. This is not uncommon. Even in real world applications such as MRI, simulated databases are used to tune image retrieval systems, see for instance [7]. Alternatively, we can imagine the retrieval of such a test set for a specific application using phantoms and their noisy acquisitions. Ideally, we would like to consider a very rich database (i.e. $N \gg 1$) in order to get a more robust estimation of the parameters, thus dealing with a very large set of constraints (1.2) that would need to be solved in each iteration of an optimisation algorithm applied to (1.1)–(1.2). The computational solution of such an optimisation problem renders expensive and therefore challenging due to the large-scale nature of the problem (1.1)–(1.2) and due to the nonsmooth nature of each constraint.

In order to deal with such large-scale problems various stochastic optimisation approaches have been presented in literature. They are based on the common idea of solving not *all* the constraints, but just a sample of them, whose size varies according to the approach one intends to use. In this paper we focus on a stochastic approximation method proposed by Byrd et al. [3,4] called *dynamic sample size* method. The main idea of this method is to consider an initial, small, training sample of the dictionary to start the algorithm with and *dynamically* increasing its size, if needed, throughout the different steps of the optimisation process. The criterion to decide whether or not the sample size has to be increased is a check on the sample variance estimates on the batch gradient. The desired trade-off between efficiency and accuracy is then obtained by starting with a small sample and gradually increasing its size till reaching the requested level of accuracy. Let us mention that the method of Byrd et al. is one among various stochastic optimisation methods, compare for instance [2,6,8–10].

Our work extends the work of [4] in two directions: firstly, in [4] the linearity of the solution map is required which is not fulfilled for our problem (1.1)–(1.2). We are going to show that the strategy of Byrd et al. can be modified for nonlinear

solution maps as the one we are considering. Secondly, in [4] the optimization algorithm is of gradient-descent type. Using a BFGS method to solve (1.1)–(1.2) we extend their approach incorporating also *second order* information in form of an approximation of the Hessian by evaluations of the sample gradient in the iterations of the optimisation algorithm.

Organisation of the paper. In the following Sect. 2 we present the Dynamic Sampling algorithm adapted to the nonlinear framework of problem (1.1)–(1.2), specifying the variance condition on the batch gradient used in our optimisation algorithm. In Sect. 3 we present the numerical results obtained for the estimation of the optimal parameters in the case of single and mixed noise estimation for the model (1.1)–(1.2) showing significant improvements in efficiency.

Preliminaries. We denote the vector of parameters we aim to estimate by $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_d) \in \mathbb{R}^d_{\geqslant 0}$. We also define by $\mathcal{S}$ the solution map that, for each constraint $k = 1, \dots, N$ of (1.2), associates to $\boldsymbol{\lambda}$ and to the noisy image f_k the corresponding Total Variation denoised solution $\hat{u}_k$, that is $\mathcal{S}(\boldsymbol{\lambda}, f_k) = \hat{u}_k$. Let us then define the reduced cost functional $J(\boldsymbol{\lambda})$ as

$$J(\boldsymbol{\lambda}) := \frac{1}{2N} \sum_{k=1}^{N} \|\mathcal{S}(\boldsymbol{\lambda}, f_k) - u_k\|^2_{L^2(\Omega)} \,. \tag{1.3}$$

We also define:

$$l(\boldsymbol{\lambda}, f_k) := \|\mathcal{S}(\boldsymbol{\lambda}, f_k) - u_k\|^2_{L^2(\Omega)} \,, \quad k = 1, \dots, N \tag{1.4}$$

as the *loss functions* of the functional J defined in (1.3) for each $k = 1, \dots, N$. For every sample $S \subset \{1, \dots, N\}$ of the database, we introduce the batch objective function:

$$J_S(\boldsymbol{\lambda}) := \frac{1}{2|S|} \sum_{k \in S} l(\boldsymbol{\lambda}, f_k). \tag{1.5}$$

2 Dynamic Sampling Schemes for Solving (1.1)–(1.2)

To design the optimisation algorithm solving (1.1)–(1.2) we follow the approach used in [5]. There, a quasi-Newton method (namely, the Broyden-Fletcher - Goldfarb-Shanno algorithm BFGS) is considered together with an Armijo backtracking linesearch rule. We combine such algorithm with a modified version of the Dynamic Sampling algorithm presented in [4, Sect. 3]. In order to compare our algorithm with the Newton-Conjugate Gradient method presented in [4, Sect. 5], we highlight that in our optimisation algorithm the Hessian matrix is never computed, but approximated efficiently by the BFGS matrix.

Our algorithm starts by selecting from the whole dataset a sample S whose size $|S|$ is small compared to the original size N. In the following iterations, if the approximation computed produces an improvement in the cost functional J, then the sample size is kept unchanged and the optimisation process continues

selecting in the next iteration a new sample of the same size. Otherwise, if the approximation computed is not a good one, a new, larger, sample size is selected and a new sample S of this new size is used to compute the new step. By starting with small sample sizes it is hoped that in the early stages of the algorithm the solution can be computed efficiently in each iteration. The key point in this procedure is clearly the rule that checks throughout the progression of the algorithm, whether the approximation we are performing is good enough, i.e. the sample size is big enough, or has to be increased. Because of this systematic check on the quality of approximation in each step of the algorithm, such sampling strategy is called *dynamic*.

As in [4], we consider a condition on the batch gradient ∇J_S which imposes at every stage of the optimisation that the direction $-\nabla J_S$ is a descent direction for J at $\boldsymbol{\lambda}$ if the following condition holds:

$$\|\nabla J_S(\boldsymbol{\lambda}) - \nabla J(\boldsymbol{\lambda})\|_{L^2(\Omega)} \leq \theta \, \|\nabla J_S(\boldsymbol{\lambda})\|_{L^2(\Omega)}, \quad \theta \in [0,1). \tag{2.6}$$

The computation of ∇J may be very expensive for applications involving large databases and nonlinear constraints, so we rewrite (2.6) as an estimate of the variance of the random vector $\nabla J_S(\boldsymbol{\lambda})$. In order to do that, recalling definitions (1.4) and (1.5) we observe that

$$\nabla J_S(\boldsymbol{\lambda}) = \frac{1}{2|S|} \sum_{k \in S} \nabla l(\boldsymbol{\lambda}, f_k). \tag{2.7}$$

We can compute (2.7) in correspondence to an optimal solution $\hat{\boldsymbol{\lambda}}$ by using [5, Remark 3.4] where a characterisation of ∇J is given in terms of the adjoint states p_k (see Sect. 3 for details). By linearity and extending to the multiple-constrained case, we get:

$$\nabla J_S(\hat{\boldsymbol{\lambda}}) = \sum_{k \in S} \sum_{i=1}^{d} \int_{\Omega} \phi_i'(\hat{u}_k, f_k) \, p_k \, dx. \tag{2.8}$$

Thanks to this characterisation, we now extend the dynamic sampling algorithm in [4] to the case where the solution map $\mathcal{S}$ is nonlinear: by taking (2.8) into account and following [4, Sect. 3] we can rewrite (2.6) as a condition on the variance of the batch gradient that reads as

$$\frac{\|Var_{k \in S}(\nabla l(\boldsymbol{\lambda}, f_k))\|_{L^1(\Omega)}}{|S|} \frac{N - |S|}{N - 1} \leq \theta^2 \, \|\nabla J_S(\boldsymbol{\lambda})\|^2_{L^2(\Omega)}. \tag{2.9}$$

For a detailed derivation of (2.9), see [4]. Condition (2.9) is the responsible for possible changes in the sample size in the optimisation algorithm and has to be checked in every iteration. If inequality (2.9) is not satisfied, a larger sample $\hat{S}$ whose size satisfies the descent condition (2.9) needs to be considered. By assuming that the change in the sample size is gradual enough such that, for any given $\boldsymbol{\lambda}$:

$$\begin{aligned} &\left\|Var_{k \in \hat{S}}(\nabla l(\boldsymbol{\lambda}, f_k))\right\|_{L^1(\Omega)} \approx \|Var_{k \in S}(\nabla l(\boldsymbol{\lambda}, f_k))\|_{L^1(\Omega)}, \\ &\left\|\nabla J_{\hat{S}}(\boldsymbol{\lambda})\right\|_{L^2(\Omega)} \approx \|\nabla J_S(\boldsymbol{\lambda})\|_{L^2(\Omega)}, \end{aligned}$$

we see that condition (2.9) is satisfied whenever we choose $|\hat{S}|$ such that

$$|\hat{S}| \geq \left\lceil \frac{N - \|Var_{k\in S}(\nabla l(\boldsymbol{\lambda}, f_k))\|_{L^1(\Omega)}}{\|Var_{k\in S}(\nabla l(\boldsymbol{\lambda}, f_k))\|_{L^1(\Omega)} + \theta^2(N-1)\,\|\nabla J_S(\boldsymbol{\lambda})\|^2_{L^2(\Omega)}} \right\rceil . \tag{2.10}$$

Conditions (2.9) and (2.10) are the key points in the optimisation algorithm we are going to present: by checking the former, one can control whether the sampling approximation is accurate enough and if this is not the case at any stage of the algorithm, by imposing the latter a new larger sample size is determined.

We remark that these two conditions force the direction $-\nabla J_S$ to be a descent direction. Steepest descent methods are known to be slowly convergent. Algorithms incorporating information coming from the Hessian are generally more efficient. However, normally the computation of the Hessian is very expensive, so Hessian-approximating methods are commonly used. In [4] a Newton-CG method is employed. There, an approximation of the Hessian matrix $\nabla^2 J_S$ is computed only on a subsample H of S such that $|H| \ll |S|$. As the sample S is dynamically changing, the subsample H will change as well (with a fixed, constant ratio) and the computation of the new conjugate gradient direction can be performed efficiently. In this work, in order to compute an approximation of the Hessian we consider the well-known BFGS method which has been extensively used in the last years because of its efficiency and low computational costs.

Before giving a full description of the resulting algorithm solving (1.1)–(1.2), we briefly comment on the linesearch rule that is employed in the update of the BFGS matrix. We choose an Armijo backtracking line search rule with curvature verification: the BFGS matrix is updated only if the curvature condition is satisfied. The Armijo criterion is:

$$J_S(\boldsymbol{\lambda_k} + \alpha_k d_k) - J_S(\boldsymbol{\lambda_k}) \leq \alpha_k \eta \nabla J_S(\boldsymbol{\lambda_k})^\top d_k \tag{2.11}$$

where the value η will be specified in Sect. 3, d_k is the descent direction of the quasi-Newton step, α_k is the length of the quasi-Newton step and $\nabla J_S(\boldsymbol{\lambda_k})$ is defined in (2.7). The positivity of the parameters is always preserved along the iterations.

We present now the BFGS optimisation with Dynamic sampling for solving (1.1)–(1.2): compared to [4, Algorithm 5.2] we stress once more that the gain in efficiency is obtained thanks to the use of BFGS instead of the Newton-CG sampling method.

3 Numerical Results

In this section we present the numerical results of the Dynamic Sampling Algorithm 1 applied to compute the numerical solution of (1.1)–(1.2). In our numerical computations we fix the parameter values as follows:

Algorithm 1. Dynamic Sampling BFGS for solving (1.1)-(1.2)

Initialize: $\boldsymbol{\lambda_0}$, sample $\mathcal{S}_0$ with $|S_0| \ll N$ and model parameter θ, $k = 0$.
while BFGS not converging, $k \geq 0$
  sample $|S_k|$ PDE constraints to solve;
  update of the BFGS matrix;
  compute direction d_k by BFGS and steplength α_k by Armijo cond. (2.11);
  define new iterate: $\boldsymbol{\lambda_{k+1}} = \boldsymbol{\lambda_k} + \alpha_k d_k$;
  if condition (2.9) **then**
    maintain the sample size: $|S_{k+1}| = |S_k|$;
  else augment S_k such that condition (2.10) is verified.
end

- We consider images of size 150×150. We approximate the differential operators by discretising with finite difference schemes with mesh step size $h = 1/$ (number of pixels in the x-direction). We use forward difference for the discretisation of the divergence operator and backward differences for the gradient. The Laplace operator is discretised by using the usual five point formula.
- The TV constraints in (1.2) are solved by means of SemiSmooth Newton (SSN) algorithms whose form depends on the ϕ's in (1.2) (cf. [5, Sect. 4]) solving regularised problems which stop if either the difference between two consecutive iterates is small enough or if the maximum number of iterations $\mathsf{maxiter} = 35$ is reached.
- In the Armijo condition (2.11) the value η is chosen to be $\gamma = 10^{-4}$.

Single noise estimation. As a toy example, we start by considering the case when the noise in the images is normally distributed. In (1.1)–(1.2), this reflects in the estimation of just one parameter λ that weights the fidelity term $\phi(u, f_k) = (u - f_k)^2$ in each constraint. Considering the training database $\{(u_k, f_k)\}_{k=1,\dots,N}$ of clean and noisy images, the problem reduces to:

$$\min_{\lambda \geq 0} \frac{1}{2N} \sum_{k=1}^{N} \|\hat{u}_k - u_k\|^2_{L^2(\Omega)} \tag{3.1}$$

where, for each k, $\hat{u}_k$ is the solution of the regularised PDE

$$-\varepsilon \Delta \hat{u}_k - \operatorname{div}\Big(h_\gamma(\nabla \hat{u}_k)\Big) + \lambda(\hat{u}_k - f_k) = 0, \quad k = 1, \dots, N. \tag{3.2}$$

In (3.2) h_γ arises from a Huber-type regularisation of the subdifferential of $|D\hat{u}_k|$ with parameter $\gamma \gg 1$ and the ε term is an artificial diffusion term that sets up the problem in the Hilbert space $H^1_0(\Omega)$ (see [5, Sect. 3] for details).

As shown in [5, Theorem 3.5] the adjoint states p_k can be computed for each constraint as the solution of the following equation

$$\begin{aligned} \varepsilon(Dp_k, Dv)_{L^2} + (h'_\gamma(D\hat{u}_k)^* Dp_k, Dv)_{L^2} \\ + \int_\Omega \lambda\ p_k\ v\ dx = -(\hat{u}_k - f_k, v)_{L^2}, \quad \forall v \in H^1_0(\Omega). \end{aligned} \tag{3.3}$$

Recalling also Eqs. (2.7)–(2.8) needed for the computation of the gradient, we can now apply Algorithm 1 to solve (3.1)–(3.2).

For the following numerical tests, the parameters of this model are chosen as follows: $\varepsilon = 10^{-12}, \gamma = 100$. The noise in the images has distribution $\mathcal{N}(0, 0.05)$. The parameter θ of the Algorithm 1, is chosen to be $\theta = 0.5$. We will comment on the sensitivity of the method to θ later on.

Table 1 shows the numerical value of the optimal parameter $\hat{\lambda}$ when varying the size of the dictionary. We measure the efficiency of the algorithms used in terms of the number of nonlinear PDEs solved during the BFGS optimisation and we compare the efficiency of solving (3.1)–(3.2) without and with the Dynamic Sampling strategy. We observe a clear improvement in efficiency when using Dynamic Sampling: the number of PDEs solved in the optimisation process is very much reduced. We note that this corresponds to an increasing number of BFGS iterations which does not appear to be an issue as BFGS iterations are themselves very fast. For the sake of computational efficiency, what really matters is the number of PDEs that need to be solved in *each* iteration of BFGS. Moreover, thanks to modern parallel computing methods and to the decoupled nature of the constraints in each BFGS iteration, solving such a reduced amount of PDEs makes the computational efforts very reasonable. In fact, we note that the size of the sample is generally maintained very small in comparison to N or just slightly increased. Computing also the relative error between the optimal parameter computed by solving all the PDEs and the one computed with

Table 1. N is the size of the database, $\hat{\lambda}$ is the optimal parameter for (3.1)–(3.2) obtained by solving all the N constraints, whereas $\hat{\lambda}_S$ is the one computed by solving the problem with Algorithm 1. The initial size S_0 is chosen to be $|S_0| = 20\,\%N$. $|S_{end}|$ of the sample at the end of the optimisation algorithm. The efficiency of the algorithms is measured in terms of the PDEs solved. We compare the accuracy of the approximation in terms of the difference $\left\|\hat{\lambda}_S - \hat{\lambda}\right\|_1 / \|\lambda_S|\|_1$.

N	$\hat{\lambda}$	$\hat{\lambda}_S$	$\lvert S_0\rvert$	$\lvert S_{end}\rvert$	Eff.	Eff. Dyn.S.	BFGS its.	BFGS its. Dyn.S.	Diff.
10	3334.5	3427.7	2	3	140	**84**	7	21	2.7 %
20	3437.0	3475.1	4	4	240	**120**	7	15	1.1 %
30	3436.5	3478.2	6	6	420	**180**	7	15	1.2 %
40	3431.5	3358.3	8	9	560	**272**	7	16	2.1 %
50	3425.8	3306.4	10	10	700	**220**	7	11	3.5 %
60	3426.0	3543.4	12	12	840	**264**	7	11	3.3 %
70	3419.7	3457.7	14	14	980	**336**	7	12	1.1 %
80	3418.1	3379.3	16	16	1120	**480**	7	15	<1 %
90	3416.6	3353.5	18	18	1260	**648**	7	18	2.3 %
100	3413.6	3479.0	20	20	1400	**520**	7	13	1.9 %

Dynamic Sampling method, we note a good level of accuracy: the difference between the two values remains below 5 %.

Figure 1 shows an example of database of brain images[1] together with the optimal denoised version obtained by Algorithm 1 for single Gaussian noise estimation.

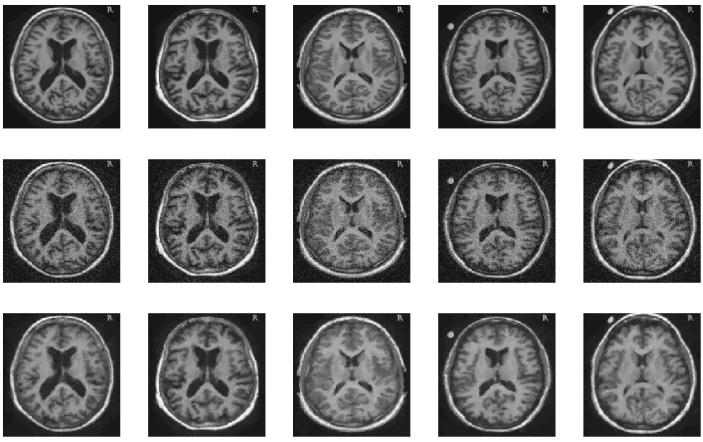


Fig. 1. Sample of 5 images of a MRI brain database: original images (upper row), noisy images (middle row) and optimal denoised images (bottom row), $\hat{\lambda_S} = 3280.5$.

Multiple noise estimation. We consider now a more interesting application of (1.1)–(1.2) where the image is corrupted by noises with different distributions. We consider the case where a combination of Gaussian and impulse noise is present. The fidelity term for the impulse distributed component is $\phi_1(u, f_k) = |u - f_k|$, whereas, as above, for the Gaussian noise we consider the fidelity $\phi_2(u, f_k) = (u - f_k)^2$, for every k. Each fidelity term is weighted by a parameter $\lambda_i, i = 1, 2$. Thus, we aim to solve:

$$\min_{(\lambda_1,\lambda_2),\ \lambda_i \geq 0} \frac{1}{2N} \sum_{k=1}^{N} \|\hat{u}_k - u_k\|^2_{L^2(\Omega)} \tag{3.4}$$

where, for each k, $\hat{u}_k$ is now the solution of the regularised PDE:

$$-\varepsilon \Delta \hat{u}_k - \mathrm{div}(h_\gamma(\nabla \hat{u}_k)) + \lambda_1 h^1_\gamma(\hat{u}_k - f_k) + \lambda_2(\hat{u}_k - f_k) = 0, \quad k = 1, \ldots, N. \tag{3.5}$$

In (3.5) the first and the second terms are as before while the third one corresponds to the Huber-type regularisation of $\mathrm{sgn}(\hat{u}_k - f_k)$. The adjoint state is computed as in [5], in a similar manner as (3.3). By taking also into account equations (2.7)–(2.8), we solve (3.4)–(3.5) with $\varepsilon = 10^{-12}, \gamma = 100$ by means of Algorithm 1.

[1] OASIS online database.

We take as example slices of the brain database shown in Fig. 1 corrupted with both Gaussian noise distributed as $\mathcal{N}(0, 0.005)$ and impulse noise with fraction of missing pixels $d = 5\,\%$, and again solve (1.1)–(1.2) by solving the PDE constraints all at once and by using Dynamic Sampling for different N. In Table 2 we report the results for the estimation of λ_1 and λ_2.

Table 2. $\hat{\lambda}_{1S}$ and $\hat{\lambda}_{2S}$ are the optimal weights for (3.4)–(3.5) estimated with Dynamic Sampling. We observe again a clear improvement in efficiency (i.e. number of PDEs solved). As above, $|S_0| = 20\,\%N$ and $\theta = 0.5$.

N	$\hat{\lambda}_{1S}$	$\hat{\lambda}_{2S}$	$\lvert S_0\rvert$	$\lvert S_{end}\rvert$	Eff.	Eff. Dyn.S.	Diff.
10	86.31	28.43	2	7	180	**70**	5.2 %
20	90.61	26.96	4	6	920	**180**	5.3 %
30	94.36	29.04	6	7	2100	**314**	5.6 %
40	88.88	31.56	8	8	880	**496**	1.2 %
50	88.92	29.81	10	10	2200	**560**	<1 %
60	89.64	28.36	12	12	1920	**336**	1.9 %
70	86.09	28.09	14	14	2940	**532**	3.3 %
80	87.68	29.97	16	16	3520	**448**	<1 %

Convergence and sensitivity. Figure 2 shows two features of Algorithm 1 applied to solve problem (3.1)–(3.2). On the left we represent the evolution of the cost functional along the BFGS iterations. Because of the sampling strategy, in the early iterations of BFGS the problem considered varies quite a lot, thus showing oscillations. Once evolving the process, the convergence is superlinear. On the right we represent the sensitivity with respect to the accuracy parameter θ (cf. (2.9)): smaller values of θ penalise larger variances on ∇J_S, thus favouring larger samples. Larger values of θ allow larger variances on ∇J_S and,

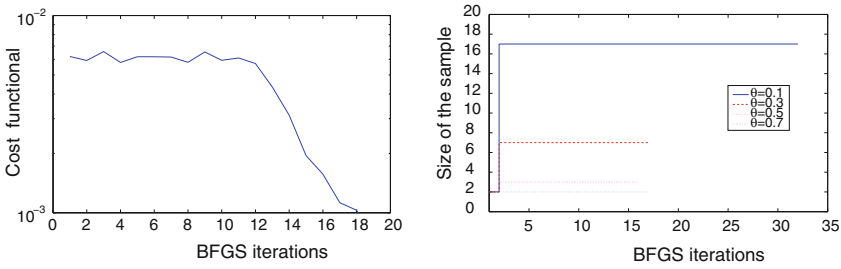


Fig. 2. *Left*: evolution of BFGS with Dynamic Sampling along the iterations. *Right*: samples size changes in Algorithm 1 for different values of θ. For each value of θ, the result is plotted till convergence. For this example $N = 20, |S_0| = 2$.

consequently, smaller sample sizes. In this case, efficiency improves, but accuracy might suffer as shown in Table 3.

Table 3. As θ increases we observe improvements upon the efficiency as smaller samples are allowed. However, the relative difference with the value estimated without sampling shows that accuracy suffers.

θ	Efficiency	Difference
0.1	516	0.07 %
0.3	246	4.3 %
0.5	92	5.9 %
0.7	68	15 %

4 Conclusions

In this paper, we propose an efficient and competitive technique to solve numerically the constrained optimisation problem (1.1)–(1.2) designed for learning the noise model in a TV denoising framework accounting for different types of noise. The set of nonsmooth PDE constraints resembles a large-size training database of clean and noisy images that allows a more robust estimation of parameters. To solve the problem, we use *Dynamic Sampling* methods, proposed in [4] for *linear* constrained problems. The idea consists in selecting just a small sample of the PDEs that need to be solved over the whole database and then, during the progression of the algorithm, verify whether such a size produces approximations that are accurate enough. Extended to our nonlinear framework, the results show a remarkable improvement in efficiency, which reflects in reduced computational times for both single noise estimations as well as for mixed ones. Further directions for future research are an accurate analysis of convergence properties of such a scheme as well as the design of a similar algorithm for the case of a L^1-regularisation on the parameter vector.

References

1. Ambrosio, L., Fusco, N., Pallara, D.: Functions of Bounded Variation and Free Discontinuity problems. In: Oxford Mathematical Monographs, p. xviii. Clarendon Press, Oxford (2000)
2. Bottou, L., Bousquet, O.: The tradeoffs of large scale learning. Adv. Neural Inf. Process. Syst. **20**, 161–168 (2008)
3. Byrd, R.H., Chin, G.M., Neveitt, W., Nocedal, J.: On the use of stochastic Hessian information in unconstrained optimisation. SIAM J. Optim. **21**(3), 977–995 (2011)
4. Byrd, R.H., Chin, G.M., Nocedal, J., Wu, Y.: Sample size selection in optimization methods for machine learning. Math. Program. Ser. B **134**, 127–155 (2012)

5. De Los Reyes, J.C., Schönlieb, C.-B.: Image denoising: learning noise distribution via nonsmooth PDE-constrained optimisation. Inverse Probl. Imaging **7**(4), 1183–1214 (2013)
6. Haber, E., Chung, M., Herrmann, F.: An effective method for parameter estimation with PDE constraints with multiple right-hand sides. SIAM J. Optim. **22**(3), 739–757 (2012)
7. Mantle, M.D., Sederman, A.J., Gladden, L.F.: Single- and two- phase flow in fixed-bed reactors: MRI flow visualisation and lattice-Boltzmann simulations. Chem. Eng. Sci. **56**, 523–529 (2001)
8. Nesterov, Y.: Primal-dual subgradient methods for convex problems. Math. Program. **120**(1), 221–259 (2009)
9. Polyak, B., Juditsky, A.: Acceleration of stochastic approximation by averaging. SIAM J. Control Optim. **30**, 838–855 (1992)
10. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. **22**(3), 400–407 (1951)

Exponential Convergence to Equilibrium for Nonlinear Reaction-Diffusion Systems Arising in Reversible Chemistry

Laurent Desvillettes[1] and Klemens Fellner [2(✉)]

[1] CMLA, ENS Cachan and CNRS, PRES UniverSud,
61, Avenue du Président Wilson, 94235 Cachan Cedex, France
desville@cmla.ens-cachan.fr
[2] Institute of Mathematics and Scientific Computing, University of Graz,
NAWI Graz Heinrichstr. 36, 8010 Graz, Austria
klemens.fellner@uni-graz.at

Abstract. We consider a prototypical nonlinear reaction-diffusion system arising in reversible chemistry. Based on recent existence results of global weak and classical solutions derived from entropy-decay related a-priori estimates and duality methods, we prove exponential convergence of these solutions towards equilibrium with explicit rates in all space dimensions.

The key step of the proof establishes an entropy entropy-dissipation estimate, which relies only on natural a-priori estimates provided by mass-conservation laws and the decay of an entropy functional.

Keywords: Reaction-diffusion equations · Entropy method · Duality method · Large-time behaviour · Convergence to equilibrium

1 Introduction

Reaction-Diffusion Systems for Reversible Chemistry

The evolution of a mixture of diffusive species $\mathcal{A}_i, i = 1, 2, \ldots, q$, undergoing a reversible reaction of the type

$$\alpha_1\mathcal{A}_1 + \cdots + \alpha_q\mathcal{A}_q \rightleftharpoons \beta_1\mathcal{A}_1 + \cdots + \beta_q\mathcal{A}_q, \quad \alpha_i, \beta_i \in \mathbb{N},$$

is modelled using mass-action kinetics (see e.g. [3–5,9] for a derivation from basic principles) in the following way:

$$\partial_t a_i - d_i\,\Delta_x a_i = (\beta_i - \alpha_i)\left(l\prod_{j=1}^{q} a_j^{\alpha_j} - k\prod_{j=1}^{q} a_j^{\beta_j}\right), \tag{1}$$

Laurent Desvillettes - LD acknowledges that the research leading to this paper was partially funded by the french "ANR blanche" project Kibord: ANR-13-BS01-0004.
Klemens Fellner - KF gratefully acknowledges the partial support of NAWI Graz.

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 96–104, 2014.
DOI: 10.1007/978-3-662-45504-3_9

where $a_i := a_i(t,x) \geq 0$ denotes the concentration at time t and point x of the species A_i and $d_i > 0$ are positive and constant diffusion coefficients.

We suppose that $x \in \Omega$, where Ω is a bounded domain of $\mathbb{R}^N$ ($N \geq 1$) with sufficiently smooth (e.g. $C^{2+\alpha}$, $\alpha > 0$) boundary $\partial\Omega$, and complement system (1) by homogeneous Neumann boundary conditions:

$$n(x)\cdot\nabla_x a_i(t,x) = 0\,, \qquad \forall t \geq 0,\ x \in \partial\Omega, \tag{2}$$

where $n(x)$ is the outer normal unit vector at point x of $\partial\Omega$.

The particular case $\mathcal{A}_1 + \mathcal{A}_2 \rightleftharpoons \mathcal{A}_3 + \mathcal{A}_4$ (that is, when $q = 4$ with $\alpha_1 = \alpha_2 = 1$, $\beta_3 = \beta_4 = 1$, $\alpha_3 = \alpha_4 = 0$ and $\beta_1 = \beta_2 = 0$) has lately received a lot of attention as a prototypical model system featuring quadratic nonlinearities, see e.g. [7,12,17]. For the sake of readability, we shall set $l = 1 = k$ (the general case can be treated without any additional difficulty) and assume that Ω is normalised (i.e. $|\Omega| = 1$). We then consider the particular case of system (1), which writes as

$$\begin{cases} \partial_t a_1 - d_1\,\Delta_x a_1 = a_3\,a_4 - a_1\,a_2, \\ \partial_t a_2 - d_2\,\Delta_x a_2 = a_3\,a_4 - a_1\,a_2, \\ \partial_t a_3 - d_3\,\Delta_x a_3 = a_1\,a_2 - a_3\,a_4, \\ \partial_t a_4 - d_4\,\Delta_x a_4 = a_1\,a_2 - a_3\,a_4, \end{cases} \tag{3}$$

together with the homogeneous Neumann boundary conditions (2).

It was first proven by Goudon and Vasseur in [17] based on an intricate use of De Giorgi's method that whenever d_1, d_2, d_3, $d_4 > 0$, there exists a global smooth solution for dimensions $N = 1, 2$. For higher space dimensions the existence of classical solutions constitutes an open problem, for which the Hausdorff dimension of possible singularities was characterised in [17]. The (technical) criticality of quadratic nonlinearities was underlined by Caputo and Vasseur in [8], where smooth solutions were shown to exist in any dimension for systems with a nonlinearity of power law type which is strictly subquadratic, see also e.g. [1].

A further related result by Hollis and Morgan [20] showed that if blow-up (here that is a concentration phenomena since the total mass is conserved) occurs in one concentration $a_i(t,x)$ at some time t and position x, then at least one more concentration has to blow-up (i.e. concentrate) at the same time and position. A proof of these results is based on a duality argument.

In [12], a duality argument in terms of entropy density variables was used to prove in an elegant way the existence of global L^2-weak solutions in any space dimension. Recently in [7], a nice improvement of the duality methods allows to show global classical solutions in 2D of the prototypical system (3)–(2) in a significantly shorter and less technical way than via De Giorgi's method.

In the present work, we shall show that exponential convergence (with explicit rates) towards the unique constant equilibrium still holds for any dimension N (see Theorem 1 below) when one considers L^2-weak solutions. The proof of Theorem 1 is based on an approach, where a quantitative entropy entropy-dissipation estimate is established, which uses only natural a-priori bounds of the system, and thus significantly improves the results of [11] and related previous results like [10,15,16,18].

The paper is organized as follows: We start in Sect. 2 by presenting a-priori bounds for our system and by overviewing the available analytical tools. Next, in Sect. 3, we prove Theorem 1 stating exponential convergence to equilibrium.

2 A Priori Estimates and Analytical Tools

2.1 Mass Conservation Laws

The conservation of the number of atoms implies (at first for all smooth solutions $(a_i)_{i=1,..,4}$ of (3) with Neumann condition (2)) that for all $t \geq 0$,

$$\begin{cases} M_{13} := \int_\Omega (a_1(t,x) + a_3(t,x))\, dx = \int_\Omega (a_1(0,x) + a_3(0,x))\, dx, \\ M_{14} := \int_\Omega (a_1(t,x) + a_4(t,x))\, dx = \int_\Omega (a_1(0,x) + a_4(0,x))\, dx, \\ M_{23} := \int_\Omega (a_2(t,x) + a_3(t,x))\, dx = \int_\Omega (a_2(0,x) + a_3(0,x))\, dx, \\ M_{24} := \int_\Omega (a_2(t,x) + a_4(t,x))\, dx = \int_\Omega (a_2(0,x) + a_4(0,x))\, dx. \end{cases} \tag{4}$$

Note that only three of the above four conservation laws are linearly independent.

2.2 Entropy Functional and Entropy Dissipation

A second set of a-priori estimates stems from the nonnegative entropy (free energy) functional $E((a_i)_{i=1,..,4})$ and the entropy dissipation $D((a_i)_{i=1,..,4}) = -\frac{d}{dt} E((a_i)_{i=1,..,4})$ associated to (3):

$$E(a_i(t,x)_{i=1,..,4}) = \sum_{i=1}^{4} \int_\Omega \Big(a_i(t,x) \log(a_i(t,x)) - a_i(t,x) + 1 \Big)\, dx, \tag{5}$$

$$\begin{aligned} D(a_i(t,x)_{i=1,..,4}) = & \sum_{i=1}^{4} \int_\Omega 4\, d_i\, |\nabla_x \sqrt{a_i(t,x)}|^2\, dx \\ & + \int_\Omega (a_1\, a_2 - a_3\, a_4) \log \left(\frac{a_1\, a_2}{a_3\, a_4} \right) (t,x)\, dx. \end{aligned} \tag{6}$$

It is easy to verify that the following entropy dissipation law holds (still for sufficiently regular solutions $(a_i)_{i=1,..,4}$ of (3) with (2)) for all $t \geq 0$

$$E(a_i(t,x)_{i=1,..,4}) + \int_0^t D(a_i(s,x)_{i=1,..,4})\, ds = E(a_i(0,x)_{i=1,..,4})\,. \tag{7}$$

The entropy decay estimate (7) implies as a first a-priori estimate that

$$a_i \in L^\infty([0,+\infty[; L\log L(\Omega)), \qquad \forall i = 1,..,4\,. \tag{8}$$

Considering in (7) that the time integral of the entropy dissipation (6) is uniformly bounded-in-time, its first component provides the estimate

$$\sqrt{a_i} \in L^2([0,+\infty[; H^1(\Omega)), \qquad \forall i = 1,..,4, \tag{9}$$

Finally, the second component of the time integral of the entropy dissipation (6) ensures that, provided that $a_3\,a_4 \in L^1_{loc}([0,+\infty[\times\bar{\Omega})$, then also $a_1\,a_2 \in L^1_{loc}([0,+\infty[\times\bar{\Omega})$. This comes out of the following classical inequality (cf. [14]), which holds for any $\kappa > 1$,

$$a_1\,a_2 \le \kappa\,a_3\,a_4 + \frac{1}{\log\kappa}\,(a_1\,a_2 - a_3\,a_4)\,\log\left(\frac{a_1\,a_2}{a_3\,a_4}\right). \tag{10}$$

Note that by letting κ be as large as necessary, this inequality also allows to prove that an approximating sequence $a_1^n\,a_2^n$ is (locally in time) weakly compact in L^1 if the sequence $a_3^n\,a_4^n$ is also weakly compact in L^1 (and when estimate (7) holds uniformly with respect to n).

Remark 1. *We remark (see [12]), that as a consequence of the first two entropy related a-priori estimates (8)–(9), global classical solutions of system (3)–(2) can be constructed only in 1D. In 2D, global L^2-weak solutions can be deduced by using Trudinger's inequality. In any higher space dimension, renormalised solution can be obtained from all three a-priori estimate (8)–(10).*

2.3 Entropy Structure and Duality Methods

The system (3)–(2) can also be rewritten in terms of the entropy density variables $z_i := a_i\log(a_i) - a_i$. By introducing the sum $z := \sum_{i=1}^4 z_i$, it holds that

$$\begin{cases} \partial_t z - \Delta_x\,(A\,z) \le 0, \qquad n(x)\cdot\nabla_x z_i(t,x) = 0, \\ A(t,x) := \frac{\sum_{i=1}^4 d_i\,z_i}{\sum_{i=1}^4 z_i} \in \left[\min\limits_{i=1,..,4}\{d_i\}, \max\limits_{i=1,..,4}\{d_i\}\right], \end{cases} \tag{11}$$

Then, by a duality argument (see e.g. [12,20,21] and the references therein), the parabolic problem (11) satisfies for all $T > 0$ and $\Omega_T = (0,T)\times\Omega$ and for all space dimensions $N \ge 1$ the following a-priori estimate

$$\|z_i\|_{L^2(\Omega_T)} \le C(1+T)^{1/2}\left\|\sum_{i=1}^4 a_{i0}(\log(a_{i0}) - 1)\right\|_{L^2(\Omega)}, \qquad i = 1,..,4, \tag{12}$$

where C is a constant independent of T, see [7,12]. Thus, given $(a_{i0})_{i=1,..,4} \in L^2(\log L)^2(\Omega)$, we have $(a_i)_{i=1,..,4} \in L^2(\log L)^2(\Omega_T)$ and the quadratic nonlinearities on the right hand side of (3) are uniformly integrable, which allows to prove the existence of global L^2-weak solutions in all space dimensions $N \ge 1$ [12]. Moreover, in 2D and in higher space dimension under the assumption of sufficiently "similar" diffusion coefficients (i.e. $\max\{d_i\} - \min\{d_i\}$ is sufficiently small), an improved duality estimate allows to show global classical solutions [7].

2.4 Equilibrium

We observe that when all the diffusivity constants $(d_i)_{i=1,..,4} > 0$ are positive, there exists a unique constant equilibrium state $(a_{i,\infty})_{i=1,..,4}$ (for which the entropy dissipation vanishes). It is defined by the unique positive constants balancing the reversible reaction $a_{1,\infty}\, a_{2,\infty} = a_{3,\infty}\, a_{4,\infty}$ and satisfying the conservation laws $a_{j,\infty} + a_{k,\infty} = M_{jk}$ for $(j,k) \in (\{1,2\},\{3,4\})$, that is:

$$\begin{cases} a_{1,\infty} = \frac{M_{13} M_{14}}{M}\,, & a_{3,\infty} = M_{13} - \frac{M_{13} M_{14}}{M} = \frac{M_{13} M_{23}}{M}\,, \\ a_{2,\infty} = \frac{M_{23} M_{24}}{M}\,, & a_{4,\infty} = M_{14} - \frac{M_{13} M_{14}}{M} = \frac{M_{14} M_{24}}{M}\,, \end{cases} \tag{13}$$

where M denotes the total initial mass $M = M_{13} + M_{24} = M_{14} + M_{23}$.

2.5 Logarithmic Sobolev Inequality

Finally, we introduce a lemma which is known to hold, but somehow without reference. We therefore follow an argument of Strook [22], which shows that Sobolev and Poincaré inequality imply the logarithmic Sobolev inequality without confining potential on a bounded domain.

Lemma 1 (Logarithmic Sobolev inequality on bounded domains). *Let Ω be a bounded domain in $\mathbb{R}^N$ such that the Poincaré (-Wirtinger) and Sobolev inequalities*

$$\|\phi - \textstyle\int_\Omega \phi\, dx\|^2_{L^2(\Omega)} \le P(\Omega)\, \|\nabla_x \phi\|^2_{L^2(\Omega)}\,, \tag{14}$$

$$\|\phi\|^2_{L^q(\Omega)} \le C_1(\Omega)\, \|\nabla_x \phi\|^2_{L^2(\Omega)} + C_2(\Omega)\, \|\phi\|^2_{L^2(\Omega)}\,, \qquad \tfrac{1}{q} = \tfrac{1}{2} - \tfrac{1}{N}\,, \tag{15}$$

hold. Then, the logarithmic Sobolev inequality

$$\int_\Omega \phi^2 \log\left(\frac{\phi^2}{\|\phi\|_2^2}\right) dx \le L(\Omega, N)\, \|\nabla_x \phi\|^2_{L^2(\Omega)} \tag{16}$$

holds (for some constant $L(\Omega, N) > 0$).

Proof (of Lemma 1). Assume firstly that $\|\phi\|_2^2 = 1$. Then, using Jensen's inequality for the measure $\phi^2\, dx$, we estimate

$$\begin{aligned} \int_\Omega \phi^2\, \log(\phi^2)\, dx &= \frac{2}{q-2} \int_\Omega \log\left(\phi^{q-2}\right)\, (\phi^2 dx) \le \frac{2}{q-2} \log\left(\int_\Omega \phi^q\, dx\right) \\ &= \frac{q}{q-2} \log\left(\|\phi\|_q^2\right) \le \frac{q}{q-2}\left(\|\phi\|_q^2 - 1\right), \end{aligned}$$

using the elementary inequality $\log x \le x - 1$. Hence, we have for general ϕ,

$$\begin{aligned} \int_\Omega \phi^2 \log\left(\frac{\phi^2}{\|\phi\|_2^2}\right) dx \le& \frac{q}{q-2}\, \left(\|\phi\|_q^2 - \|\phi\|_2^2\right) \\ \le& \frac{q}{q-2}\, C_1\, \|\nabla_x \phi\|_2^2 + \frac{q}{q-2}\, (C_2 - 1)\, \|\phi\|_2^2, \end{aligned}$$

using the Sobolev inequality (15). Now, in case when $\int_\Omega \phi\, dx = 0$, inequality (16) follows directly from Poincaré inequality (14). Otherwise, considering $\tilde{\phi} = \phi - \int_\Omega \phi\, dx$, a lengthy calculation [13] shows that

$$\int_\Omega \phi^2 \log\left(\frac{\phi^2}{\|\phi\|_2^2}\right) dx \le \int_\Omega \tilde{\phi}^2 \log\left(\frac{\tilde{\phi}^2}{\|\tilde{\phi}\|_2^2}\right) dx + 2\,\|\tilde{\phi}\|_2^2\,,$$

and the inequaltiy (16) follows from Poincaré inequality (14).

Remark 2. *On convex domains Ω, an alternative proof of (16) consists in building a limiting procedure with a sequence of logarithmic Sobolev inequalities on $\mathbb{R}^N$ (see e.g. [2, 6]) with a convex confining potential, which is made constant inside the bounded domain (by using the Holley-Strook perturbation lemma [19]) and tends to infinity outside of the bounded domain.*

3 Exponential Convergence to Equilibrium via the Entropy Method

In this section, we prove exponential convergence towards equilibrium (with explicit rates) for weak solutions of system (3) (and thus also for classical solution whenever they are known to exist) in all space dimensions $N \ge 1$:

Theorem 1. *Let Ω be a bounded domain with sufficiently smooth boundary (e.g. $\partial\Omega \in C^{2+\alpha}$, $\alpha > 0$) such that Lemma 1 holds. Let $(d_i)_{i=1,..,4} > 0$ be positive diffusion coefficients. Let the initial data $(a_{i,0})_{i=1,..,4}$ be nonnegative functions of $L^2\,(\log L)^2(\Omega)$ with positive masses $(M_{jk})_{(j,k)\in(\{1,2\},\{3,4\})} > 0$ (see (4)). Then, the global solution a_i of (3)–(2) (weak or classical as shown to exist in [7, 12]) decay exponentially towards the positive equilibrium state $(a_{i,\infty})_{i=1,..,4} > 0$ defined by (13):*

$$\sum_{i=1}^4 \|a_i(t,\cdot) - a_{i,\infty}\|^2_{L^1(\Omega)} \le C_1\Big(E((a_{i,0})_{i=1,..,4}) - E((a_{i,\infty})_{i=1,..,4})\Big)\, e^{-C_2\, t},$$

for all $t \ge 0$ and for constants C_1 and C_2, which can be explicitly computed.

Remark 3. *The above Theorem generalises to all space dimensions the convergence result obtained in [11]. It avoids a slowly growing L^∞-bound (available only in 1D and maybe 2D) by using the logarithmic Sobolev inequality (16) to control the relative entropy of the concentrations a_i w.r.t. their spatial averages $\overline{a_i} = \int_\Omega a_i\, dx$ (recall that $|\Omega| = 1$), which themself are controlled by the mass conservation laws (4). The remaining part of the proof follows then from [11].*

Note also that exponential decay towards equilibrium in $L^p(\Omega)$ with $1 < p < 2$ follows by interpolation the $L^2(\Omega)$-bounds (12).

Proof (of Theorem 1). The proof is based on an entropy method, where the entropy dissipation $D((a_i)_{i=1,..,4}) = -\frac{d}{dt}E((a_i)_{i=1,..,4}) = -\frac{d}{dt}(E((a_i)_{i=1,..,4}) - E((a_{i,\infty})_{i=1,..,4}))$ is controlled from below in terms of the relative entropy with respect to equilibrium. That is, we look for an estimate like

$$\begin{aligned} D((a_i)_{i=1,..,4}) &\geq C\left(E((a_i)_{i=1,..,4}) - E((a_{i,\infty})_{i=1,..,4})\right) \\ &= C\sum_{i=1}^{4}\int_\Omega \left[a_i \log\left(\frac{a_i}{a_{i,\infty}}\right) - (a_i - a_{i,\infty})\right] dx, \end{aligned} \tag{17}$$

for a constant C provided that all the conservation laws (4) are observed. Then, a simple Gronwall lemma yields exponential convergence in relative entropy to the equilibrium $(a_{i,\infty})_{i=1,..,4}$. Furthermore, convergence in L^1 as stated in Theorem 1 follows from a Cziszar-Kullback type inequality [11, Proposition 4.1].

In order to establish the entropy-entropy dissipation estimate (17), we firstly split the relative entropy

$$\begin{aligned} E((a_i)_{i=1,..,4}) - E((a_{i,\infty})_{i=1,..,4}) &= E((a_i)_{i=1,..,4}) - E((\overline{a_i})_{i=1,..,4}) \\ &\quad + E((\overline{a_i})_{i=1,..,4}) - E((a_{i,\infty})_{i=1,..,4}), \end{aligned}$$

into – roughly speaking – the relative entropy of the concentrations a_i w.r.t. their averages $\overline{a_i}$ and the relative entropy of the averages $\overline{a_i}$ w.r.t. the equilibrium $a_{i,\infty}$.

The first term can be estimated thanks to the logarithmic Sobolev inequality (16) (recall the conservation laws (4)) by

$$\begin{aligned} E((a_i)_{i=1,..,4}) - E((\overline{a_i})_{i=1,..,4}) &= \sum_{i=1}^{4}\int_\Omega a_i \log\left(\frac{a_i}{\overline{a_i}}\right) dx \\ &\leq L(\Omega)\sum_{i=1}^{4}\int_\Omega |\nabla_x \sqrt{a_i}|^2\, dx, \end{aligned}$$

which is clearly bounded by the entropy dissipation $D((a_i)_{i=1,..,4})$ in (6).

On the other hand, estimating the second relative entropy can be done in the following way: We define

$$\phi(x,y) = \frac{x\,\ln(x/y) - (x-y)}{(\sqrt{x}-\sqrt{y})^2} = \phi(x/y, 1),$$

which is a continuous function on $(0,\infty)\times(0,\infty)$. Note that thanks to the conservation laws (4), we have $\phi(\overline{a_i}/a_{i,\infty}, 1) \leq C(M)$. We can then write

$$\begin{aligned} E((\overline{a_i})_{i=1,..,4}) - E((a_{i,\infty})_{i=1,..,4}) &= \sum_{i=1}^{4}\left[\overline{a_i}\log\left(\frac{\overline{a_i}}{a_{i,\infty}}\right) - (\overline{a_i} - a_{i,\infty})\right] \\ &\leq \sum_{i=1}^{4}\phi(\overline{a_i}, a_{i,\infty})\left|\sqrt{\overline{a_i}} - \sqrt{a_{i,\infty}}\right|^2 \leq C(M)\sum_{i=1}^{4}\left|\sqrt{\overline{a_i}} - \sqrt{a_{i,\infty}}\right|^2. \end{aligned}$$

Finally, the expression $\sum_{i=1}^4 \left|\sqrt{a_i} - \sqrt{a_{i,\infty}}\right|^2$ is bounded in terms of equation (47) in [11, Lemma 3.2], which itself is bounded by the entropy dissipation $D((a_i)_{i=1,..,4})$ in (6) with a constant, which can be explicitly estimated. This finishes the proof of the entropy entropy-dissipation estimate (17), which implies explicit exponential convergence to equilibrium in relative entropy.

The proof of Theorem 1 follows then by recalling the Cziszar-Kullback type inequality [11, Proposition 4.1].

References

1. Amann, H.: Global existence for semilinear parabolic problems. J. Reine Angew. Math. **360**, 47–83 (1985)
2. Arnold, A., Markowich, P., Toscani, G., Unterreiter, A.: On convex Sobolev inequalities and the rate of convergence to equilibrium for Fokker-Planck type equations. Comm. Partial Diff. Equ. **26**(1–2), 43–100 (2001)
3. Bisi, M., Desvillettes, L.: From reactive Boltzmann equations to reaction-diffusion systems. J. Stat. Phys. **125**(1), 249–280 (2006)
4. Bisi, M., Conforto, F., Desvillettes, L.: Quasi-steady-state approximation for reaction-diffusion equations. Bull. Inst. Math., Acad. Sin. (N.S.) **2**, 823–850 (2007)
5. Bothe, D., Pierre, M.: Quasi-steady-state approximation for a reaction-diffusion system with fast intermediate. J. Math. Anal. Appl. **368**(1), 120–132 (2010)
6. Carrillo, J., Jüngel, A., Markowich, P., Toscani, G., Unterreiter, A.: Entropy dissipation methods for degenerate parabolic problems and generalized Sobolev inequalities. Monatsh. Math. **133**(1), 1–82 (2001)
7. Cañizo, J.A., Desvillettes, L., Fellner, K.: Improved duality estimates and applications to reaction-diffusion equations. Commun. Partial Diff. Equ. **39**(6), 1185–1204 (2014)
8. Caputo, C., Vasseur, A.: Global regularity of solutions to systems of reaction-diffusion with sub-quadratic growth in any dimension. Commun. Partial Diff. Equ. **34**(10–12), 1228–1250 (2009)
9. De Masi, A., Presutti, E.: Mathematical Methods for Hydrodynamic Limits. Lecture Notes in Mathematics, vol. 1501. Springer, Heidelberg (1991)
10. Desvillettes, L., Fellner, K.: Exponential decay toward equilibrium via entropy methods for reaction-diffusion equations. J. Math. Anal. Appl. **319**(1), 157–176 (2006)
11. Desvillettes, L., Fellner, K.: Entropy methods for reaction-diffusion equations: slowly growing a priori bounds. Revista Matemática Iberoamericana **24**(2), 407–431 (2008)
12. Desvillettes, L., Fellner, K., Pierre, M., Vovelle, J.: About global existence for quadratic systems of reaction-diffusion. J. Adv. Nonlinear Stud. **7**(3), 491–511 (2007)
13. Deuschel, J.-D., Stroock, D.W.: Large Deviations. Pure and Applied Mathematics, vol. 137. Academic Press Inc., Boston (1989)
14. Di Perna, R.J., Lions, P.L.: On the Cauchy problem for Boltzmann equations: global existence and weak stability. Ann. Math. **130**, 321–366 (1989)
15. Glitzky, A., Gröger, K., Hünlich, R.: Free energy and dissipation rate for reaction-diffusion processes of electrically charged species. Appl. Anal. **60**(3–4), 201–217 (1996)

16. Glitzky, A., Hünlich, R.: Energetic estimates and asymptotics for electro-reaction-diffusion systems. Z. Angew. Math. Mech. **77**, 823–832 (1997)
17. Goudon, T., Vasseur, A.: Regularity analysis for systems of reaction-diffusion equations. Ann. Sci. Ec. Norm. Super. **43**(1), 117–141 (2010)
18. Gröger, K.: Free energy estimates and asymptotic behaviour of reaction-diffusion processes. Preprint 20, WIAS, Berlin (1992)
19. Holley, R., Stroock, D.: Logarithmic Sobolev inequalities and stochastic Ising models. J. Statist. Phys. **46**(5–6), 1159–1194 (1987)
20. Hollis, S.L., Morgan, J.J.: On the blow-up of solution to some semilinear and quasilinear reaction-diffusion systems. Rocky Mt. J. Math. **24**(4), 1447–1465 (1994)
21. Pierre, M., Schmitt, D.: Blowup in reaction-diffusion systems with dissipation of mass. SIAM Rev. **42**(1), 93–106 (2000)
22. Stroock, D.: Logarithmic Sobolev inequalities for gibbs states. Lect. Notes Math. **1563**, 194–228 (1993)

A High-Order Semi-Lagrangian/Finite Volume Scheme for Hamilton-Jacobi-Isaacs Equations

Maurizio Falcone[1(✉)] and Dante Kalise[2]

[1] Department of Mathematics, Sapienza - University of Rome,
P.le Aldo Moro 2, 00185 Rome, Italy
falcone@mat.uniroma1.it
[2] Johann Radon Institute for Computational and Applied Mathematics,
Altenberger StraßE 69, 4040 Linz, Austria
dante.kalise@oeaw.ac.at

Abstract. We present a numerical scheme for the approximation of Hamilton-Jacobi-Isaacs equations related to optimal control problems and differential games. In the first case, the Hamiltonian is convex with respect to the gradient of the solution, whereas the second case corresponds to a non convex (minmax) operator. We introduce a scheme based on the combination of semi-Lagrangian time discretization with a high-order finite volume spatial reconstruction. The high-order character of the scheme provides an efficient way towards accurate approximations with coarse grids. We assess the performance of the scheme with a set of problems arising in minimum time optimal control and pursuit-evasion games.

Keywords: Hamilton-Jacobi-Isaacs equations · High-order schemes · Semi-Lagrangian schemes · Finite volume methods · Optimal control · Differential games

1 Introduction

The numerical approximation of Hamilton-Jacobi-Isaacs (henceforth HJI) equations appears as a crucial step in many fields of applications, including optimal control, image processing, fluid dynamics, robotics and geophysics. In general, these equations do not have regular solutions even if the data and the coefficients are regular, and therefore many efforts have been devoted to the development and the analysis of approximation schemes for such problems. The convergence of the schemes is understood in the sense of viscosity solutions; it is well known (see e.g. [5,20]) that viscosity solutions are typically Lipschitz continuous, and therefore the main difficulty is to have a good resolution around the singularities, and a good accuracy in the parts of the domain where the solution is regular.

This research was supported by the following grants: AFOSR Grant no. FA9550-10-1-0029, ITN-Marie Curie Grant no. 264735-SADCO, and the FWF-START project *Sparse Approximation and Optimization in High Dimensions*.

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 105–117, 2014.
DOI: 10.1007/978-3-662-45504-3_10

The theory of approximation schemes for viscosity solutions has been developed starting from the huge literature existing for the numerical solution of conservation laws in one dimension, exploiting the relation between entropy solutions and viscosity solutions. More precisely, the viscosity solution can be written as the space integral of the corresponding entropy solution and this relation can be also applied to the construction of numerical schemes, by simply integrating in space the schemes for conservation laws. At the very beginning, these techniques were successfully applied to the study of the class of monotone schemes; in this framework, the rate of convergence is limited to first order. Some of these schemes, like finite differences for instance, are used over structured grids and are strictly related to the above mentioned methods for conservation laws. Other approximation schemes, like the Finite Volume Method and semi-Lagrangian schemes, can easily work on unstructured grids and are based on different ideas, e.g. on the Hopf-Lax representation formula. In all these cases, the role of monotonicity is important to guarantee the convergence to the viscosity solution, and a general result for monotone schemes applied to second order fully nonlinear equations has been proved by Barles and Souganidis in [6]. Although a complete list of the contributions to numerical methods for HJI equations goes beyond the scopes of this paper, let us quote the application of Godunov/central schemes [1,2], antidissipative and SuperBee/UltraBee [10,11], MUSCL [26], and WENO schemes [12,29].

A natural way to overcome the limitations of monotone schemes is by the application of high-order approximations. For a given accuracy, these methods can achieve acceptable error levels in coarser grids, with a considerably reduced number of nodes in comparison with low-order, monotone schemes. This can be a crucial point when the dimension of the problem is high or when complex computations are required at every grid node; both situations naturally arise in the context of HJI equations stemming from optimal control and differential games. In this paper we propose the coupling between a semi-Lagrangian (SL) time discretization with a finite volume reconstruction in space. High-order SL schemes for HJI equations have been first considered for a semi-discretization in time in [18], and for the fully discrete scheme in [19]. A convergence analysis based on the condition $\Delta x = O(\Delta t^2)$ is carried out in [21]. The adaptation of the theory to weighted ENO reconstructions is presented in [14], along with a number of numerical tests comparing the various high-order versions of the scheme. Other numerical tests, mostly in higher dimension and concerned with applications to front propagation and optimal control, are presented in [13]. Let us mention that a first convergence result for a class of motonone Finite Volume schemes has been proved in [27].

The paper is organized as follows.

In Sect. 2, we illustrate our ideas with a setting related to minimum time optimal control and differential games, leading to stationary HJI equation. In Sect. 3, we deal with a high-order approximation scheme based on a coupling between a semi-Lagrangian discretization in time and a Finite Volume spatial

reconstruction. Finally, in Sect. 4 we present some numerical experiments assessing the performance and accuracy of the proposed scheme.

2 HJI Equations Arising in Optimal Control and Differential Games

As we mentioned in the introduction, HJI equations often arise in optimal control and differential games; whenever a feedback controller is sought, the application of the Dynamic Programming Principle (DPP) leads to HJI equations, which can be time-dependent or stationary. Among a wide class of problems, in this section we illustrate our ideas by means of minimum time optimal control and pursuit-evasion games.

Let us start by considering consider system dynamics of the form

$$\begin{cases} \dot{y}(t) = f(y, \alpha(t)) \quad \text{for } t > 0\,, \\ y(0) = x\,, \end{cases} \tag{1}$$

where $y \in \mathbb{R}^n$ is the state, $\alpha : [0, +\infty) \to A$ is the control and $f : \mathbb{R}^n \times A \to \mathbb{R}^n$ is the controlled vector field. To get a unique trajectory for every initial condition and a given control function, we will always assume that f is continuous with respect to both variables, and Lipschitz continuous with respect to the state space (uniformly in α). Moreover, we will assume that the controls are measurable functions of time so that we can apply the Carathéodory theorem for the Cauchy problem (1).

In the minimum time optimal control problem, we want to minimize the time of arrival to a given target $\mathcal{T}$. The cost will be given by

$$t(x, \alpha) := \begin{cases} \inf\{t : y_x(t; \alpha) \in \mathcal{T}\} & \\ +\infty & \text{if } y_x(t; \alpha) \notin \mathcal{T}\ \forall t. \end{cases} \tag{2}$$

By the application of the DPP one can prove that the minimum time function

$$T(x) := \inf_{\alpha \in \mathcal{A}} t(x, \alpha)$$

satisfies the Bellman equation

$$\max_{a \in A} \{-f(x, a) \cdot DT\} = 1\,,$$

in the domain where T is finite (the so-called reachable set). Introducing the change of variable

$$v(x) := \begin{cases} 1/\mu & \text{if } t_x(a, b) = +\infty\,, \\ 1/\mu(1 - e^{-\mu t_x(a,b)}) & \text{elsewhere}\,, \end{cases} \tag{3}$$

where μ is a free positive parameter to be suitably chosen, one can characterize T as the unique viscosity solution of the following Dirichlet problem

$$\begin{cases} \mu v(x) + \max_{a \in A}\{-f(x,a)\cdot Dv\} = 1 & \text{for } x \in \mathbb{R}^n \setminus \mathcal{T}\,, \\ v(x) = 0 & \text{for } x \in \partial\mathcal{T}\,. \end{cases} \tag{4}$$

Another interesting example comes from the DPP approximation of the Hamilton-Jacobi-Isaacs equations related to *pursuit-evasion games* (see [5] for more details). Player-a (the *pursuer*) wants to catch player-b (the *evader*) who is escaping, and the controlled dynamics for each player are known. To simplify the notations, we will denote by $y(t) = (y_P(t), y_E(t))$ the state of the system, where $y_P(t)$ and $y_E(t)$ are the positions at time t of the pursuer and of the evader, both belonging to $\mathbb{R}^n$, and by $f : \mathbb{R}^{2n} \times A \times B \to \mathbb{R}^{2n}$ the dynamics of the system (clearly, here the dynamics depend on the controls for both players). The payoff is defined as the time of capture but, in order to have a fair game, we need to restrict the strategies of the players to the so-called *non-anticipating strategies* (i.e. they cannot exploit the knowledge of the future strategy of the opponent). These strategies will be denoted respectively by α and β. Given the strategies $\alpha(\cdot)$ and $\beta(\cdot)$ for the first and the second player, we can define the corresponding time of capture as

$$t_x(\alpha[\beta],\beta) = \inf\ \left\{t > 0 :\ y_P(t) = y_E(t)\right\}.$$

If there is no capture for those strategies we set $t_x(\alpha[\beta],\beta) = +\infty$. Then we can define the lower time of capture as

$$T(x) = \inf_{\alpha\in\mathcal{A}} \sup_{\beta\in\mathcal{B}} t_x(\alpha[\beta],\beta),$$

and again T can be infinite if there is no way to catch the evader from the initial position of the system x. In order to get a fixed point problem and to deal with finite values, it is useful to scale time by the change of variable (3), which corresponds to the payoff

$$J_x(\alpha,\beta) = \int_0^{t_x(\alpha,\beta)} e^{-\mu t}dt$$

The rescaled minimal time will be given by

$$v(x) = \inf_{\alpha\in\mathcal{A}} \sup_{b\in\mathcal{B}} J_x(\alpha[\beta],\beta).$$

Assuming v to be continuous, the application of the DPP leads to the following characterization of the value function

$$\begin{cases} v + \min_{b\in B}\max_{a\in A}\{-Dv\cdot f(x,a,b)\} = 1 \text{ on } \mathbb{R}^n \setminus \mathcal{T}, \\ v(x) = 0 \qquad\qquad\qquad\qquad\qquad\quad \text{on } \partial\mathcal{T}. \end{cases} \tag{5}$$

Note that the equation is complemented by the natural homogeneous boundary condition on the target $T(x) = v(x) = 0$. If $v(\,\cdot\,)$ is continuous, then v is a viscosity solution in $\mathbb{R}^n \setminus \mathcal{T}$ of the Dirichlet problem (5).

3 Semi-Lagrangian Schemes for HJI Equations

In this section we introduce the main building blocks for the construction of semi-Lagrangian/finite volume schemes for HJI equations of the form (4)–(5). The general procedure is decomposed into a time discretization step, and a space discretization procedure. In the time discretization step, the system dynamics (1) are approximated by a suitable integration rule, and the DPP is applied on its discrete-time version. In the space discretization procedure, the resulting HJI equation is then approximated over a finite set of elements. The same procedure holds for all the problems presented in the previous section, however, for the sake of simplicity, this section is illustrated by means of the minimum time problem and its associated HJB Eq. (4) (which is also a particular case of (5) for a single player setting).

Time Discretization

For the implementation of a time disretization procedure, we follow the ideas presented in [18]. The first step towards the construction of a high-order scheme for the equations (4)–(5) is to consider discrete time approximation of the system dynamics (1) of the form

$$\begin{aligned} y_{n+1} &= y_n + h\,\Phi(y_n, A_n, h)\,, \qquad \text{for } n > 0\,, \\ y_0 &= x\,, \end{aligned} \tag{6}$$

where $h > 0$ corresponds to a time discretization parameter, $\Phi = \Phi(y_n, A_n, h)$ is the Henrici function of a one-step approximation of the dynamical system, and A_n stands for a multidimensional control defined accordingly to the order q of the numerical integrator,

$$A_n = (a_n^0, a_n^1, \ldots, a_n^q)\,, \quad A_n \in A^{q+1}\,, q \geq 0.$$

Particular cases of the aforementioned setting are

(i) Explicit Euler's method: $\Phi(y_n, A_n, h) = f(y_n, a_n)$, with $A_n = a_n^0$.
(ii) Midpoint rule: $\Phi(y_n, A_n, h) = f(y_n + hf(y_n, a_n^0)/2, a_n^1)$, with $A_n = (a_n^0, a_n^1)$.
(iii) Fourth-order Runge-Kutta scheme:

$$\Phi(y_n, A_n, h) = \frac{1}{6}(K_0 + 2K_1 + 2K_2 + K_3)\,, \quad A_n = (a_n^0, a_n^1, a_n^2, a_n^3)\,,$$
$$K_0 = f(y_n, a_n^0)\,, K_1 = f(y_n + h\frac{K_0}{2}, a_n^1)\,, K_2 = f(y_n + h\frac{K_1}{2}, a_n^2)\,, K_3 = f(y_n + hK_2, a_n^3).$$

The application of the DPP for the discrete-time dynamics leads to an approximation of Eq. (4) of the form

$$\begin{cases} v_h(x) = \min\limits_{A_n \in A^q} \{\beta v_h(x + h\Phi(x, A_n, h))\} + 1 - \beta & \text{for } x \in \mathbb{R}^n \setminus \mathcal{T}\,, \\ v(x) = 0 & \text{for } x \in \partial\mathcal{T}\,, \end{cases} \tag{7}$$

where $\beta = e^{-h}$ appears as a consequence of the application of the Kruzkhov transform (3) to the discrete version of the minimum time solution (in this case with $\mu = 1$).

Note that, despite having introduced an approximation in time for Eq. (4), the resulting semi-discrete version (7) is still continuously defined over the state space. In order to implement a fully-discrete computational scheme, it is necessary to realize this expression in a bounded domain with a finite set of elements. Classical schemes for HJI equations of this form are based upon finite difference discretizations, where the domain $\Omega \subset \mathbb{R}^n$ over which the solution is sought, is discretized into a set of grid points, and the approximation is understood in a pointwise sense. A natural problem in this setting arises from the fact that the r.h.s. of Eq. (7) requires the evaluation of v_h at the *arrival points* $x+h\Phi(x, A_n, h)$, which are not necessarily part of the grid. In the low-order version of the SL scheme, this evaluation is performed via piecewise linear interpolation from the grid values, whereas in this work we focus on a high-order definition of such an operation. We follow an approach based on a Finite Volume approximation of the problem. For a given mesh parameter k, and a set of central nodes $\{x_i\}_{i=1}^N$, the domain is discretized into a set of cells $\Omega_i = [x_i - k/2, x_i + k/2]$. Instead of considering pointwise nodal values of v_h, the solution will be represented by a set of cell-averaged values $V := \{v_i\}_{i=1}^N$ defined as

$$v_i := \frac{1}{k} \int_{x_i - k/2}^{x_i + k/2} v_h(x)\, dx\,, \quad i = 1, \ldots, N.$$

Its is straightforward to see that the exact expression for the averaged values of the solution of (7) is given by

$$\begin{aligned} v_i &= T_{k,i}(v_h) \quad \text{for } i = 1, \ldots, N\,, \\ T_{k,i}(v_h) :&= \frac{1}{k} \int_{x_i - k/2}^{x_i + k/2} \left\{ \min_{A_n \in A^q} \{\beta v_h(x + h\Phi(x, A_n, h))\} + 1 - \beta \right\} dx\,, \\ v(x) &= 0 \quad \text{for } x \in \partial\mathcal{T}\,. \end{aligned} \tag{8}$$

A first approximation is introduced when the integral in (8) is replaced by a suitable Gaussian quadrature rule

$$T_{k,i}(v_h) \approx \frac{1}{k} \sum_i w_i \left\{ \min_{A_n \in A^q} \{\beta v_h(x_i + h\Phi(x_i, A_n, h))\} + 1 - \beta \right\}, \tag{9}$$

where x_i and w_i are Gauss points and weights inside the $i-$th cell, respectively. This expression requires the evaluation of the exact v_h at a set of arrival points, which is not available. Analogously to the grid-based schemes, we approximate this evaluation with an interpolation operator $I = I[V]$ defined upon the set of cell averages, i.e.

$$v_h(x_i + h\Phi(x_i, A_n, h)) \approx I[V](x_i + h\Phi(x_i, A_n, h)).$$

where $I : \mathbb{R}^N \rightarrow\!\!\!\!\rightarrow S_k$ corresponds to a WENO (weighted essentially non-oscillatory) interpolation routine performed over the averaged dataset V. The WENO reconstruction procedure and related numerical schemes date back to the work of [25], in the context of numerical methods for conservation laws, as a way of circumventing Godunov's barrier theorem by considering nonlinear (on the data) reconstruction procedures for the implementation of high-order accurate schemes. As it has been shown in [21], the use of a WENO interpolation procedure can be considered as a building block in high-order, semi-Lagrangian schemes for time-dependent HJB equations, whereas here we introduce an application to static HJI equations. We now briefly describe the main ideas for a 1D WENO reconstruction.

From a set of cell values V and a polynomial degree r, the WENO reconstruction procedure yields a set of polynomials $P = \{p_i(x)\}_{i=1}^N$ of degree r, holding standard interpolation properties

$$v_i = \frac{1}{k}\int_{\Omega_i} p_i(x)\,dx\,, \quad v(x) = p_i(x) + o(\Delta x^r)\,,\ \forall x \in \Omega_i, \tag{10}$$

and an *essentially non-oscillatory condition* [24]; in general, such an interpolant is built by considering a set of stencils per cell, and weighting them according to a smoothness indicator. Several variations of this procedure can be found in the literature; for illustration purposes, we restrict ourselves to the reconstruction procedure presented in [4], on its 1D version, and reconstruction degree 2. In this case, given a set of averaged values V, the reconstruction procedure seeks, for every cell, a local quadratic expansion upon a linear combination of Legendre polynomials rescaled in local coordinates $\xi = [-1/2,\ 1/2]$, expressed in the form

$$p(\xi) = v_0 + v_\xi p_1(\xi) + v_{\xi\xi} p_2(\xi),$$

with

$$p_1(\xi) = \xi\,, \qquad p_2(\xi) = \xi^2 - \frac{1}{12}.$$

We assign the subscript "0" to the cell where we compute the coefficients, other values indicating location and direction with respect to v_0 (note that the notation is coherent with the fact that the first coefficient in the expansion v_0, holds $v_0 = v_i$, i.e., the centered value). Next, for this particular problem we define three stencils

$$S^1 = \{v_{-2}, v_{-1}, v_0\}\,, \quad S^2 = \{v_{-1}, v_0, v_1\}\,, \quad S^3 = \{v_0, v_1, v_2\}\,,$$

and in every stencil we compute a polynomial of the form

$$v^{(i)}(\xi) = v_0^{(i)} + v_\xi^{(i)} p_1(\xi) + v_{\xi\xi}^{(i)} p_2(\xi) \qquad i = 1, 2, 3.$$

Imposing the conservation condition (10), the coefficients are given by

$$\begin{aligned}
&S^1 : v_\xi^{(1)} = -2v_{-1} + v_{-2}/2 + 3v_0/2, \quad v_{\xi\xi}^{(1)} = (v_{-2} - 2v_{-1} + v_0)/2\,,\\
&S^2 : v_\xi^{(2)} = (v_1 - v_{-1})/2, \quad v_{\xi\xi}^{(2)} = (v_{-1} - 2v_0 + v_1)/2\,,\\
&S^3 : v_\xi^{(3)} = -3v_0/2 + 2v_1 - v_2/2, \quad v_{\xi\xi}^{(3)} = (v_0 - 2v_{-1} + v_2)/2.
\end{aligned}$$

For every polynomial we calculate a smoothness indicator defined as

$$IS^{(i)} = \sum_{l=1}^{r} \int_{\Omega_0} k^{2l-1} \left(\frac{\partial^l p^{(i)}}{\partial x^l} \right)^2 dx \,,$$

where r is the polynomial reconstruction degree (in our case $r = 2$), and which in our case yields to

$$IS^{(i)} = \left(p_\xi^{(i)} \right)^2 + \frac{13}{3} \left(p_{\xi\xi}^{(i)} \right)^2 .$$

This leads to the following WENO weights:

$$\omega^{(i)} = \frac{\alpha^{(i)}}{\sum_{i=1}^{3} \alpha^{(i)}} \,, \quad \alpha^{(i)} = \frac{\lambda^{(i)}}{(\epsilon + IS^{(i)})^r} \,,$$

where ϵ is a parameter introduced in order to avoid division by zero; usually $\epsilon = 10^{-12}$. The scheme is rather insensitive to the parameter r, which we set $r = 5$. The parameter λ is usually computed in an optimal way to increase the accuracy of the reconstruction at certain points; we opt for a centered approach instead, thus $\lambda^{(1)} = \lambda^{(3)} = 1$, while $\lambda^{(2)} = 100$. Finally, the expression for the 1D reconstructed polynomial at the $i - th$ cell is given by

$$p_i(\xi) = \omega^{(1)} p^{(1)}(\xi) + \omega^{(2)} p^{(2)}(\xi) + \omega^{(3)} p^{(3)}(\xi).$$

Having defined all the buildings blocks for a fully-discrete, high-order approximation of Eq. (4), we need to solve the following nonlinear system

$$\begin{aligned} v_i &= [T_{k,i}(V)]_i \quad \text{for } i = 1, \dots, N \,, \\ [T_{k,i}(V)]_i &:= \frac{1}{k} \sum_i w_i \left\{ \min_{A_n \in A^q} \{\beta I[V](x_i + h\Phi(x_i, A_n, h))\} + 1 - \beta \right\} , \\ v(x) &= 0 \quad \text{for } x \in \mathcal{T} \,, \\ v(x) &= 1 \quad \text{for } x \in \partial\Omega^c \backslash \mathcal{T}. \end{aligned} \tag{11}$$

Note that we added an additional boundary condition related to the external part of the computational domain which is not the target, computationally equivalent to setting a high value which is neglected in the minimization procedure for the interior elements. The computational domain must be set accordingly to this condition, in order to generate a consistent result. With respect to the solution of the nonlinear system (11), the approach which we follow is motivated by the standard approach undertaken in the low-order setting, which is to solve the system by some variation of a fixed point iteration

$$V^{n+1} = T(V^n) \,, \tag{12}$$

which, in the low-order monotone scheme, is well-justified since T is a contraction mapping. A key point is the fact that the corresponding linear interpolation operator is monotone, which is lost in the high-order scheme. However, it is

still possible to develop a convergence theory for interpolation operators that are not monotone but have additional properties such as the WENO operator. In [21], a convergence framework has been developed for time-dependent HJB equations, and recently in [8], convergence results have been obtained for the stationary case. One of the advantages of this setting is the vast amount of available literature dealing with acceleration techniques for HJI iterative solvers (we refer the reader to [3] and references therein for a recent update on such methods).

4 Numerical Examples

We now present two numerical examples assessing the performance of the proposed scheme. We recall that, although we will present examples dealing with minimum time optimal control and pursuit-evasion games, the presented ideas can be applied in a straightforward manner to infinite/finite horizon optimal control, reachability analysis and differential games.

A Two-Dimensional Minimum Time Problem

We begin by considering a two-dimensional minimum time problem. In this first example, system dynamics are given by

$$f(x, y, (a_1, a_2)) = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix},$$

the domain is $\Omega =]-1, 1[^2$, the target is $\mathcal{T} = \partial\Omega$, $h = 0.8k$ and $A = \{(1,0),(0,1),(-1,0),(0,-1)\}$ is the set of 4 directions pointing to the facets of the unit square. The exact solution for this problem is the distance function to the unit square. As the characteristic curves for this problem correspond to straight lines moving towards the boundary, integration in time can be achieved exactly with a solver of any order. We consider then a Euler discretization in time, and a two-dimensional WENO reconstruction of order 2 in space. The multidimensional WENO reconstruction is based on a product of unidimensional reconstructions along every direction (we refer the reader to [4] for the specific version used in this test). Convergence rates and errors are shown in Table 1. Note that the second order of the space interpolation is achieved for the $\|\cdot\|_1$ norm, while a lower order is observed for the $\|\cdot\|_\infty$ norm. This is expected from the fact that the solution is not differentiable across the kinks of the solution. Note that if error computation is performed over a restricted zone excluding non-differentiable points, as in Table 2, higher order of accuracy and convergence are achieved for both norms, as it is generally expected for WENO-based schemes. However, this might not be the case for any high-order scheme. As an example, in Fig. 1, it can be seen that if a generic quadratic interpolation is used to build a similar scheme, spurious oscillations arise in non-differentiable areas, degenerating the high-order accuracy of the scheme. This latter justifies the use of WENO reconstruction operators in space, as they are accordingly designed in order to detect and penalize highly oscillatory stencils.

Table 1. Errors for the 2D minimum time problem with a second-order WENO reconstruction.

k	$\|\cdot\|_\infty$-error	$\|\cdot\|_\infty$-order	$\|\cdot\|_1$-error	$\|\cdot\|_1$-order	#iterations
0.08	0.0023	–	3.667e-4	–	21
0.04	0.0011	1.0641	1.068e-4	1.7791	30
0.02	5.711e-4	0.946	2.887e-5	1.8878	54
0.01	2.966e-4	0.945	7.51e-6	1.9427	104

Table 2. Error computation (as in Table 1) over a restricted zone excluding non-differentiable kinks.

k	$\|\cdot\|_\infty$-error	$\|\cdot\|_\infty$-order	$\|\cdot\|_1$-error	$\|\cdot\|_1$-order
0.08	7.763e-9	–	1.051e-9	–
0.04	5.362e-10	3.8558	1.438e-10	2.8696
0.02	3.523e-11	3.9279	1.0758e-11	3.7406
0.01	2.257e-12	3.9643	7.1681e-13	3.9076

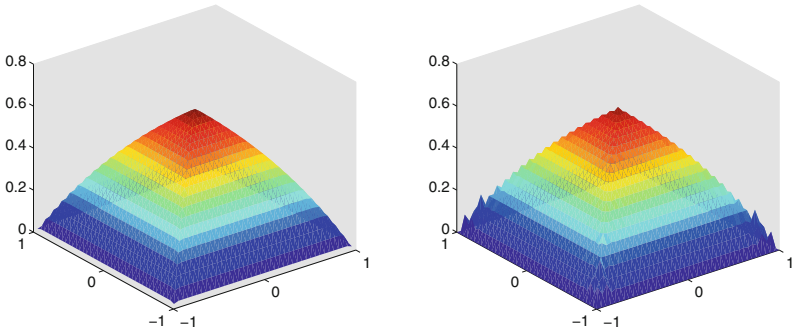


Fig. 1. 2D minimum time problem: value function for different schemes. Left: high-order scheme with WENO reconstruction in space. Right: high-order scheme using quadratic interpolation.

A Reduced-Coordinate Pursuit-Evasion Game

In a second example, we consider a 1D pursuit-evasion game with dynamics given by

$$\dot{x}_P = v_P a\,, \qquad \dot{x}_E = v_E b\,,$$

where v_P and v_E denote the velocity of the pursuer and the evader respectively; $a \in [0, 1]$ and $b \in [-1, 1]$ are control variables. By defining the reduced coordinate $x = x_E - x_p$, the game is written as

$$\dot{x} = v_E b - v_P a.$$

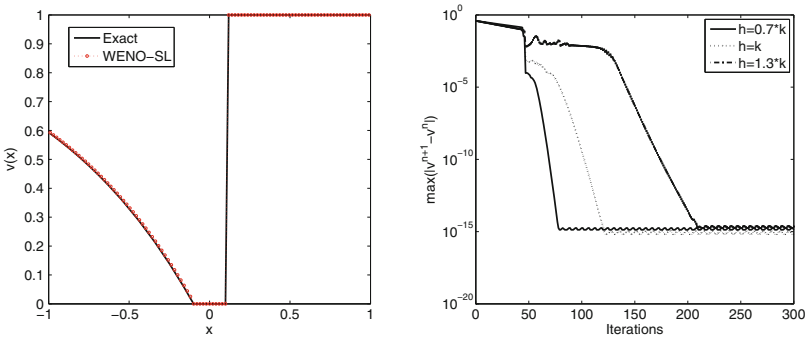


Fig. 2. SL/FV scheme for a 1D differential game. Left: exact and approximated solution for 100 elements. Right: oscillatory, but convergent behavior is achieved for the fixed point iteration (12), for different values of h.

If we consider the target set $\mathcal{T} = B(0, R)$, the exact solution is given by

$$v(x) = \begin{cases} 1 - exp(-|x + R|) & \text{if } x < R \\ 0 & \text{if } x \in (-R, R) \\ 1 & \text{if } x > R\,. \end{cases}$$

We implement our SL/FV scheme with a fourth-order RK scheme in time and a WENO reconstruction in space of degree 2; results are shown in Fig. 2. A natural advantage of high-order methods is the level of accuracy that can be reached with a reduced number of elements, which is particularly relevant when fixed point iterations of the form (12) involving *min* or *minmax* operators are considered. However, at it has been previously discussed, for high-order schemes the fixed point operator is not a contraction anymore, and convergence has to be understood in a different sense. In [8], the ϵ- monotonicity concept has been introduced in order to characterize the convergence behavior of such high-order schemes. Figure 2 illustrates this situation, as for different values of h and k, convergence of the fixed point iteration is achieved in an oscillatory way, whereas the oscillation behavior decreases when $h = h(k)$ is reduced.

5 Concluding Remarks

We have introduced a semi-Lagrangian/finite volume scheme for the approximation of HJI equations. The main building blocks are a high-order approximation of the system dynamics, combined with high order of accuracy in space via a WENO interpolation operator. The resulting fully-discrete scheme is then solved by means of a fixed point iteration. High-order of accuracy is observed in smooth regions, and the convergence of the fixed point iteration is achieved as long as in the spatial-resolution building block, non-oscillatory interpolation operators are considered. Further developments in the directions of this paper will include

the implementation of high-dimensional interpolation routines, as well as the construction of adaptive schemes with an ad-hoc refinement criterion.

References

1. Abgrall, R.: Numerical discretization of the first-order hamilton-jacobi equation on triangular meshes. Comm. Pure Appl. Math. **49**, 1339–1373 (1996)
2. Abgrall, R.: Numerical discretization of boundary conditions for first order Hamilton-Jacobi equations. SIAM J. Numer. Anal. **41**, 2233–2261 (2003)
3. Alla, A., Falcone, M., Kalise, D.: An efficient policy iteration algorithm for dynamic programming equations. ArXiv preprint: 1308.2087 (2013)
4. Balsara, D., Dumbser, M., Munz, C.-D., Rumpf, T.: Efficient, high accuracy ADER-WENO schemes for hydrodynamics and divergence-free magnetohydrodynamics. J. Comput. Phys. **228**, 2480–2516 (2009)
5. Bardi, M., Capuzzo-Dolcetta, I.: Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations. Birkhäuser, Boston (1997)
6. Barles, G., Souganidis, P.E.: Convergence of approximation schemes for fully nonlinear second order equations. Asymptotic Anal. **4**, 271–283 (1991)
7. Bauer, F., Grüne, L., Semmler, W.: Adaptive spline interpolation for Hamilton-Jacobi-Bellman equations. Appl. Numer. Math. **56**, 1196–1210 (2006)
8. Bokanowski, O., Falcone, M., Ferretti, R., Grüne, L., Kalise, D., Zidani, H.: Value iteration convergence of ϵ-monotone schemes for stationary Hamilton-Jacobi equations. Preprint 33 pp. (2014)
9. Bokanowski, O., Garcke, J., Griebel, M., Klompmaker, I.: An adaptive sparse grid semi-Lagrangian scheme for first order Hamilton-Jacobi Bellman equations. J. Sci. Comput. **55**, 575–605 (2013)
10. Bokanowski, O., Zidani, H.: Anti-dissipative schemes for advection and application to Hamilton-Jacobi-Bellman equations. J. Sci. Comput. **30**, 1–33 (2007)
11. Bryson, S., Kurganov, A., Levy, D., Petrova, G.: Semi-discrete central-upwind schemes with reduced dissipation for Hamilton-Jacobi equations. IMA J. Numer. Anal. **25**, 113–138 (2005)
12. Bryson, S., Levy, D.: Mapped WENO and weighted power ENO reconstructions in semi-discrete central schemes for Hamilton-Jacobi equations. Appl. Numer. Math. **56**, 1211–1224 (2006)
13. Carlini, E., Falcone, M., Ferretti, R.: An efficient algorithm for Hamilton-Jacobi equations in high dimension. Comput. Vis. Sci. **7**, 15–29 (2004)
14. Carlini, E., Ferretti, R., Russo, G.: A weighted essentially nonoscillatory, large time- step scheme for Hamilton-Jacobi equations. SIAM J. Sci. Comput. **27**, 1071–1091 (2005)
15. Crandall, M.G., Lions, P.L.: Two approximations of solutions of Hamilton-Jacobi equations. Math. Comp. **43**, 1–19 (1984)
16. Crandall, M.G., Majda, A.: Monotone difference approximations for scalar conservation laws. Math. Comp. **34**, 1–21 (1980)
17. Falcone, M.: Numerical methods for differential games via PDEs. Int. Game Theory Rev. **8**, 231–272 (2006)
18. Falcone, M., Ferretti, R.: Discrete time high-order schemes for viscosity solutions of Hamilton-Jacobi-Bellman equations. Numer. Math. **67**, 315–344 (1994)
19. Falcone, M., Ferretti, R.: Semi-Lagrangian schemes for Hamilton-Jacobi equations, discrete representation formulae and Godunov methods. J. Comp. Phys. **175**, 559–575 (2002)

20. Falcone, M., Ferretti, R.: Semi-Lagrangian Schemes for Linear and Hamilton-Jacobi Equations. SIAM, Philadelphia (2014)
21. Ferretti, R.: Convergence of semi-Lagrangian approximations to convex Hamilton-Jacobi equations under (very) large Courant numbers. SIAM J. Numer. Anal. **40**, 2240–2253 (2003)
22. Harten, A., Osher, S.: Uniformly high-order accurate nonoscillatory schemes. I. SIAM J. Numer. Anal. **24**, 279–309 (1987)
23. Harten, A., Osher, S., Engquist, B., Chakravarthy, S.R.: Some results on uniformly high-order accurate essentially nonoscillatory schemes. Appl. Numer. Math. **2**, 347–377 (1986)
24. Harten, A., Osher, S., Engquist, B., Chakravarthy, S.R.: Uniformly high order accurate essentially non-oscillatory schemes. III. J. Comput. Phys. **71**, 231–303 (1987)
25. Liu, X., Osher, S.: Weighted essentially non-oscillatory schemes. J. Comput. Phys. **115**, 200–212 (1994)
26. Lions, P.L., Souganidis, P.E.: Convergence of MUSCL and filtered schemes for scalar conservation laws and Hamilton-Jacobi equations. Numer. Math. **69**, 441–470 (1995)
27. Kossioris, G., Makridakis, C., Souganidis, P.E.: Finite volume schemes for Hamilton-Jacobi equations. Numer. Math. **83**, 427–442 (1999)
28. Osher, S.: Convergence of generalized MUSCL schemes. SIAM J. Numer. Anal. **22**, 947–961 (1985)
29. Zhang, Y.-T., Shu, C.-W.: High-order WENO schemes for Hamilton-Jacobi equations on triangular meshes. SIAM J. Sci. Comput. **24**, 1005–1030 (2002)

Simultaneous Material and Topology Optimization Based on Topological Derivatives

Jannis Greifenstein(✉) and Michael Stingl

Chair of Applied Mathematics 2,
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU),
Nägelsbachstr. 49b, 91052 Erlangen, Germany
{greifenstein,stingl}@math.fau.de
http://www.math.fau.de

Abstract. We use an asymptotic expansion of the compliance cost functional in linear elasticity to find the optimal material inside elliptic inclusions. We extend the proposed method to material optimization on the whole domain and compare the global quality of the solutions for different inclusion sizes. Specifically, we use an adjusted free material optimization problem, that can be solved globally, as a global lower material optimization bound. Finally, the asymptotic expansion is used as a topological derivative in a simultaneous material and topology optimization problem.

Keywords: Material optimization · Topology optimization · Material orientation · Asymptotic expansions · Discrete material optimization

1 Introduction

We investigate material and topology optimization of compliance problems in two dimensions. To this end, we first present an asymptotic formula[1] of the compliance functional for the insertion of a number of ellipsoidal bodies in an elastic domain, originally derived in [1]. Later, we study numerically the feasibility of replacing all material using the same asymptotic expansion as for the ellipses and finally make use of the formula as topological derivative.

The problem described above is by far not new. There are many publications dealing with similar types of problems. For the rotational optimization considered later, in [2] an analytical formula for the strain energy is derived, to directly compute the optimal material orientation. In [3], this has been embedded into a structural optimization algorithm for compliance minimization. A similar approach is discussed in [4] for a plate model. The method proposed in this article, however, can be used for a broader spectrum in material optimization as well,

The authors want to thank the German Research Foundation (DFG) for funding this research work within Collaborative Research Centre 814, subproject C2.

[1] We note that the asymptotic formulae are also available for the three-dimensional case.

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 118–127, 2014.
DOI: 10.1007/978-3-662-45504-3_11

such as discrete material optimization. The algorithm for simultaneous material and topology optimization presented at the end of this article is very similar to topology gradient methods, see e.g. [5]. For more references, see [6].

2 Material and Topology Optimization

We consider a domain $\Omega \subset \mathbb{R}^2$ of isotropic elastic material. The domain is subject to exterior traction and other boundary conditions (e.g. homogeneous Dirichlet conditions). The objective is to find the optimal material C^0 in a set of admissible materials $\mathcal{C}$ to insert into an inclusion, for which the compliance as defined in (1) is minimized.

The elastic body is modeled by the equation of linear elasticity

$$\int_\Omega C_{ijkl}(x)\varepsilon_{ij}(u)\varepsilon_{kl}(v)\,\mathrm{dx} = \int_\Gamma f u\,\mathrm{ds},$$

with the displacement field u, the linearized strains $\varepsilon_{ij}(u) = \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial u_i}\right)$ and a traction f. We use Voigt notation to denote the fourth-order stiffness tensor C_{ijkl} by a symmetric 3×3-matrix and the strains and stresses by a vector

$$C = \begin{pmatrix} C_{1111} & C_{1122} & \sqrt{2}C_{1112} \\ & C_{2222} & \sqrt{2}C_{2212} \\ \text{sym.} & & 2C_{1212} \end{pmatrix}, \qquad \varepsilon = \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \sqrt{2}\varepsilon_{12} \end{pmatrix}, \qquad \sigma = \begin{pmatrix} \sigma_{11} \\ \sigma_{22} \\ \sqrt{2}\sigma_{12} \end{pmatrix}.$$

The stresses are given by *Hooke's law*

$$\sigma = C\varepsilon.$$

2.1 Optimal Material in Elliptic Inclusions

We insert a finite number of inclusions ω_i, $i = 1, \ldots, n_{\text{ell}}$ with $n_{\text{ell}} > 1$ and centers $z_1, \ldots, z_{n_{\text{ell}}}$ into the domain Ω and search for the optimal material to be used in the inclusions. For an exemplary setup of boundary conditions and loads, a sketch of a possible problem specification is shown in Fig. 1. We place the elliptic inclusions on a regular grid within the domain Ω, so that the center points of the inclusions are distributed equidistantly. In the sketch in Fig. 1, we have $n_{\text{ell}} = 100$ disjoint elliptic inclusions. In the remaining domain

$$\Omega^1 := \Omega \setminus \bigcup_{i=1}^{n_{\text{ell}}} \omega_i,$$

we insert an isotropic matrix material C^1, and, in each inclusion, a material C_i, $i = 1, \ldots, n_{\text{ell}}$.

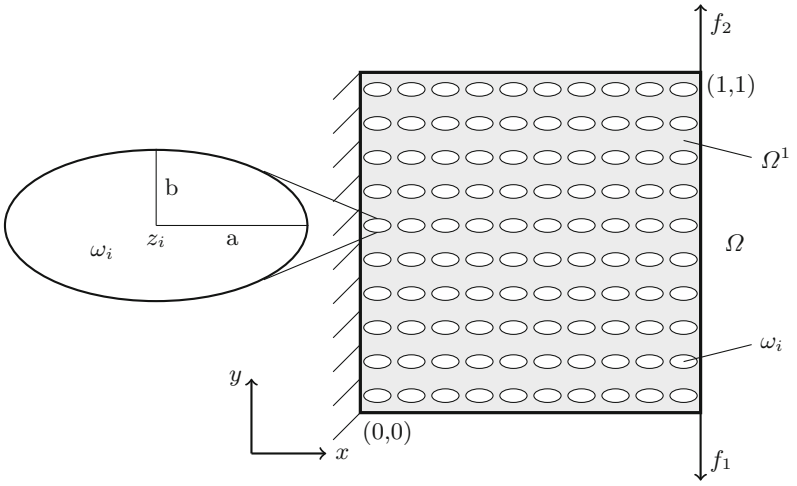


Fig. 1. Example for placing elliptic inclusions $\omega_i, i = 1, \ldots, n_{\text{ell}}$, in the domain Ω.

Now, the compliance function of the elastic body Ω (without inclusions) with respect to a set of given load functions $f_k \in [L^2(\Gamma)]^2,\ k \in \mathcal{K} = \{1, 2, \ldots, n_{\text{loads}}\}$ applied to a part Γ of $\partial\Omega$ is defined as

$$\mathcal{J}_c(f) = \sum_{k \in \mathcal{K}} \int_{\Gamma} f_k(x) u_k(x) \, \mathrm{ds}. \tag{1}$$

By virtue of the asymptotic expansion (cf. [1]) for the two-dimensional case, the compliance function $\mathcal{J}_{c(C^0)}(f)$ of a body with a single inclusion of material C^0 can be approximated as

$$\left| \mathcal{J}_{c(C^0)}(f) - \left(\mathcal{J}_c(f) - \frac{h^2}{2} \varepsilon(u;z)^\top P_{(C^0)} \varepsilon(u;z) \right) \right| \leq c h^{5/2} (1 + |\ln h|) \|f\|^2, \tag{2}$$

where h is a dimensionless scaling parameter for the elliptic inclusion ω, $\varepsilon(u;z)$ is the strain corresponding to the displacement of the domain Ω without inclusion evaluated at the center z of the ellipse, and $P_{(C^0)}$ is the so called polarization matrix given by

$$P_{(C^0)} = -|\omega| \left(C^1 - C^0 + (C^1 - C^0) \left(\mathbb{I}_3 - \Psi (C^1 - C^0) \right)^{-1} \Psi (C^1 - C^0) \right). \tag{3}$$

The matrix Ψ in formula (3) depends solely on the isotropic matrix material and is, for the stretched coordinates $\omega = \{\xi : \xi_1^2/a^2 + \xi_2^2/b^2 < 1\},\ \xi = x/h$, computed by

$$\Psi = \int_{\partial\omega} D(\nu(\xi)) \left(D(\nabla_\xi) \Phi(\xi) \right)^\top \mathrm{ds}_\xi, \qquad D(x)^\top = \begin{pmatrix} x_1 & 0 & 2^{-1/2} x_2 \\ 0 & x_2 & 2^{-1/2} x_1 \end{pmatrix}, \tag{4}$$

where ν denotes the exterior normal of the ellipse ω and a, b are the semi-major and semi-minor axes of the ellipse, respectively, cf. Fig. 1. Finally, Φ in (4) is the fundamental solution given by

$$\Phi(x) = c_1 \begin{pmatrix} -(\lambda+3\mu)\ln(r) - (\lambda+\mu)\frac{x_2^2}{r^2} & (\lambda+\mu)\frac{x_1 x_2}{r^2} \\ (\lambda+\mu)\frac{x_1 x_2}{r^2} & -(\lambda+3\mu)\ln(r) - (\lambda+\mu)\frac{x_1^2}{r^2} \end{pmatrix}, \quad (5)$$

where $c_1 = \frac{1}{4\pi}\frac{1}{\mu(\lambda+2\mu)}$, $r = \sqrt{x_1^2 + x_2^2}$ and λ and μ are the Lamé parameters corresponding to the isotropic material C^1, see e.g. [7, Chap. 3]. A more detailed explanation may be found in [1, Sect. 3.1].

In order to minimize the compliance for a single inclusion with center z_i we can now use the asymptotic expansion (2) and The inserted material will be chosen out of a set of admissible materials $\mathcal{C}$

$$\begin{aligned} \min_{C^0 \in \mathcal{C}} \mathcal{J}_{c(C^0)}(f) &\approx \min_{C^0 \in \mathcal{C}} \mathcal{J}_c(f) - \frac{h^2}{2} \varepsilon(u; z_i)^\top P_{(C^0)} \varepsilon(u; z_i) \\ &= \mathcal{J}_c(f) + \frac{h^2}{2} \min_{C^0 \in \mathcal{C}} \left(-\varepsilon(u; z_i)^\top P_{(C^0)} \varepsilon(u; z_i) \right). \end{aligned}$$

Thus, taking into account that $J_c(f)$ is independent of the rotation angle and noting that h^2 is a constant scaling parameter, the functional

$$\mathcal{D}_c(C^0, z_i) := -\varepsilon(u; z_i)^\top \, P_{(C^0)} \, \varepsilon(u; z_i) \quad (6)$$

can be used as an approximate model to find the optimal rotation angle resulting in the optimization problem

$$\min_{C^0 \in \mathcal{C}} \mathcal{D}_c(C^0, z_i). \quad (7)$$

We note that the functional $\mathcal{D}_c$ depends on the inserted material C^0 only via the polarization matrix $P_{(C^0)}$ and on the displacement field $u(x)$ only locally through the evaluation of the strain at center z_i of the ellipse. Furthermore, only a single evaluation of the state problem for the unperturbed domain is required.

Due to the local character of (6) the optimal orientation of $n_{\text{ell}} < \infty$ non-intersecting inclusions at once can be approximated by the solution of n_{ell} optimization problems of type (7) or equivalently by the solution of the problem

$$\min_{(C_0^0, \ldots, C_{n_{\text{ell}}}^0) \in \mathcal{C}^{n_{\text{ell}}}} \mathcal{J}^{\text{asymp}}_{c(C_0^0, \ldots, C_{n_{\text{ell}}}^0)} := \sum_{i=1}^{n_{\text{ell}}} \min_{C_i^0} D_c(C_i^0, z_i), \quad (8)$$

which is separable in terms of the different materials $C_0^0, \ldots, C_{n_{\text{ell}}}^0$ used in the different inclusions. The latter is certainly only an approximation of the original simultaneous material optimization problem

$$\min_{(C_0^0, \ldots, C_{n_{\text{ell}}}^0) \in \mathcal{C}^{n_{\text{ell}}}} \mathcal{J}_{c(C_0^0, \ldots, C_{n_{\text{ell}}}^0)}, \quad (9)$$

however it is shown in [6] for the optimal rotation of an orthotropic material, that the results for this approximation are close to the solution of the original problem. Moreover, we will later on compare the quality of the approximated solution to rigorous lower bounds. We want to stress that still only a single evaluation of the state problem for the unperturbed domain is required, which allows for a highly efficient numerical solution.

The optimization procedure for solving (8) is performed by the following algorithm:

(S1) choose matrix material C^1 and admissible materials for inclusions $\mathcal{C}$;
(S2) compute Ψ from (4);
(S3) define loads and boundary conditions;
(S4) solve state problem without inclusions for isotropic material;
foreach *inclusion* ω_i, $i = 1, \dots, n_{ell}$ **do**
  (S5) solve $\min_{C_i^0 \in \mathcal{C}} \mathcal{D}_c(C_i^0, z_i)$ to global optimality;
end

Algorithm 1. Basic algorithm for minimization of the compliance based on the asymptotic model.

2.2 Admissible Material Choices

In order to avoid local minima, the set of admissible materials $\mathcal{C}$ should allow for a global solution of the local material optimization problem (7) in a reasonable time. Interestingly, the easiest choice here would be a set of discrete materials. In the following, however, we concentrate on parametric material formulations. Using the properties of the asymptotic expansion, we can give a wider class of parametrizations that lead to globally solvable problems. According to [6] (Theorem 2.8, p. 14), the polarization matrix (3) is positive definite if $(C^0)^{-1} - (C^1)^{-1}$ is negative definite. If the isotropic material C^1 is then chosen s.t. every $C^0 \in \mathcal{C}$ is strictly stiffer, it can then be shown that the functional (6) is convex for any linear material parametrization. An example of this would be the so-called free material optimization (FMO, see e.g. [8,9]):

$$\mathcal{C}_{\mathrm{FMO}} := \left\{ \begin{pmatrix} e_1 & e_2 & e_3 \\ e_2 & e_4 & e_5 \\ e_3 & e_5 & e_6 \end{pmatrix} \succeq \underline{\tau} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \; e_1 + e_4 + e_6 = \overline{\rho} \right\}, \tag{10}$$

where $\underline{\tau}$ is a fixed lower eigenvalue bound and $\overline{\rho}$ bounds the total stiffness of the material tensor. The positivity constraint is still linear as a semidefinite program, s.t. the resulting problem stays convex.

Furthermore, we will now consider the usually difficult problem of rotational optimization that is also vastly simplified by the breakdown to local minimization problems. An example of this would be the variation of angle and stiffness

$$C^0(\theta, s) := \Theta(\theta)^\top C(s) \Theta(\theta), \; C(s) := \begin{pmatrix} 1.0 + s & 0.5 & 0 \\ 0.5 & 15 - s & 0 \\ 0 & 0 & 1.0 \end{pmatrix} \tag{11}$$

where we rotate an orthotropic material by an angle θ using the orthogonal rotation matrix Θ defined as

$$\Theta(\theta) = \begin{pmatrix} \cos(\theta)^2 & \sin(\theta)^2 & -\sqrt{2}/2\sin(2\theta) \\ \sin(\theta)^2 & \cos(\theta)^2 & \sqrt{2}/2\sin(2\theta) \\ \sqrt{2}/2\sin(2\theta) & -\sqrt{2}/2\sin(2\theta) & \cos(2\theta) \end{pmatrix}. \tag{12}$$

In the results section, we will consider two different admissible material sets based on this parametrization, namely

$$\mathcal{C}_{\theta,s} := \left\{ C^0(\theta, s) : \theta \in [0, \pi],\ s \in [0, 14] \right\} \tag{13}$$

and a pure rotational optimization

$$\mathcal{C}_{\theta,s=0} := \left\{ C^0(\theta, s) : \theta \in [0, 2\pi],\ s = 0 \right\}. \tag{14}$$

Now this parametrization fails the proposed linearity. However, for any fixed rotation angle θ, the local minimization problems (8) are still strictly convex. It follows, that the problem can be solved as a bilevel problem

$$\min_{\theta_i, s_i} \mathcal{J}_c(C_i^0(\theta_i, s_i)) = \min_{\theta_i} G(\theta_i)$$
$$G(\theta_i) := \min_{s_i} \mathcal{J}_{c(C^0(\theta_i, s_i))},$$

which still allows for a global solution with moderate cost as there is only a single primary variable left. Note that this could be done similarly for more complicated linear parametrizations with rotation or in 3D with 2–3 rotation angles.

Lastly, we consider an orthotropic material parametrization using engineering constants with fixed Poisson's ratio $\nu_{xy} = \nu_{yx} = 0.3$:

$$C(\theta, E_x, E_y, G_{xy}) = \Theta(\theta)^\top \begin{pmatrix} \frac{E_x}{1-0.09} & \frac{\sqrt{0.09 E_x E_y}}{1-0.09} & 0 \\ \frac{\sqrt{0.09 E_x E_y}}{1-0.09} & \frac{E_y}{1-0.09} & 0 \\ 0 & 0 & 2G_{xy} \end{pmatrix} \Theta(\theta), \tag{15}$$

with elasticity moduli E_x, E_y, shear modulus G_{xy} and Θ as in (12). We define the admissible material set corresponding to (13) as

$$\mathcal{C}_{\text{Eng}} := \left\{ C(\theta, E_x, E_y, G_{xy}) : \theta \in [0, \pi],\ E_x, E_y \in [1, 15],\ G_{xy} \in [0.5, 7.5] \right\}.$$

2.3 From Elliptic Inclusions to Material Optimization

While the asymptotic model (2) rigorously holds only for elliptic inclusions of small size, in the following we will also numerically investigate the behavior when replacing the material inside squared patches of finite elements. Choosing the elements properly, the material in the whole domain can be replaced this way with the FE patches still being disjoint. Using a large number of patches, the size of the inclusion stays small compared to the domain size. Thus, we will study increasingly bigger ellipses and compare the compliance values of the different parametrizations to global lower material optimization bounds computed with an FMO solver.

Validation Methods. For the numerical evaluation, we discretize the domain Ω using rectangular finite elements. This discretization is necessary to compute the displacements used in the asymptotic expansion. The elliptic inclusions are approximated by those finite elements, for which the coordinates of their center point are contained in the inclusion ω_i. In order to obtain the actual compliance value for the optimization result, the material used in those elements is then replaced by the optimal value of C_i^0. When replacing all material, we use equally sized squared FE patches that are uniformly distributed, disjoint and cover the whole domain.

Furthermore, we compare the compliance values to a global lower material optimization bound determined by solving a modified FMO problem. Specifically, we solve

$$\min_{(C_0^0,\dots,C_{n_{\text{ell}}}^0)\in\mathcal{C}_{\text{FMO}}^{n_{\text{ell}}}} \mathcal{J}_{c(C_0^0,\dots,C_{n_{\text{ell}}}^0)} \tag{16}$$

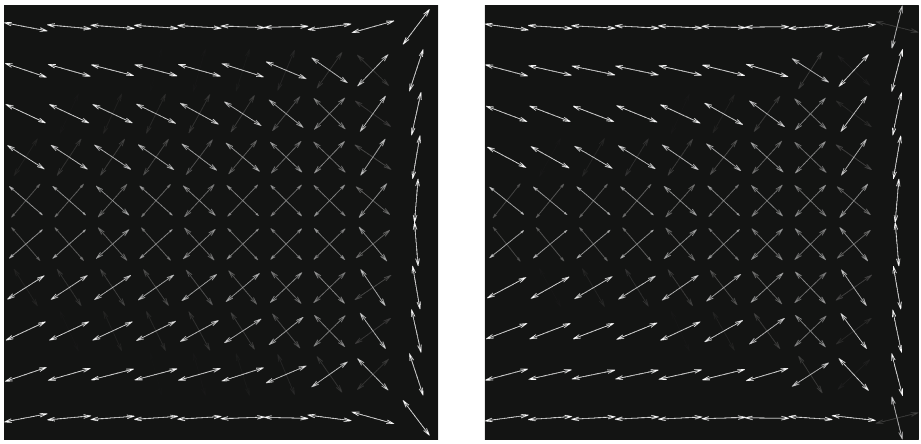
with $\mathcal{C}_{\text{FMO}}$ as in (10) and the bounds $\underline{\tau}$ and $\overline{\rho}$ chosen as close as possible to the ones used in the specific material parametrization, s.t. all possible tensors are a subset of $\mathcal{C}_{\text{FMO}}$. For a more detailed description of the method, see [6]. Note that within these bounds, any physically admissible material may be used and that this problem is convex for the compliance cost functional. Thus, the problem is solved globally using the algorithm described in [9] and we obtain a global lower compliance bound.

Numerical Results. We consider the example from Fig. 1 with 10×10 ellipses and discretize the domain Ω using 100×100 finite elements. We compare the different admissible material sets as defined in Sect. 2.2. For $\mathcal{C}_{\text{FMO}}$ we choose $\underline{\tau} = 1$ as lower eigenvalue bound and $\overline{\rho} = 17$ as upper trace bound both in the asymptotic material optimization as in the FMO solver. The results are shown in Table 1 and a visualization in Fig. 2. Although the error compared to the exact FMO result increases heavily with the size of the inclusions, this is largely due to the decrease of the overall compliance value. The absolute value does not increase much from the largest ellipses to the squared FE patches.

For the squared FE patches, we furthermore study the different parametrizations separating the domain into 50×50 patches. The results are found in Table 2 and Fig. 3(a). We can see, that for the parametrization with nonlinear

Table 1. Compliance values and FMO comparison for increasing ellipse size.

$a = b$:	0.02		0.04		0.05		squared patch	
$\mathcal{C}_{\theta,s=0}$	15.420	1.0 %	11.053	6.15 %	8.3173	14.2 %	4.3730	42.3 %
$\mathcal{C}_{\theta,s}$	15.332	0.42 %	10.738	3.12 %	7.7140	5.93 %	3.5668	16.1 %
$\mathcal{C}_{\text{Eng}}$	15.328	0.39 %	10.647	2.25 %	7.6135	4.55 %	3.5014	13.9 %
$\mathcal{C}_{\text{FMO}}$	15.289	0.14 %	10.551	1.33 %	7.4840	2.77 %	3.2912	7.10 %
FMO	15.268		10.413		7.2821		3.0730	

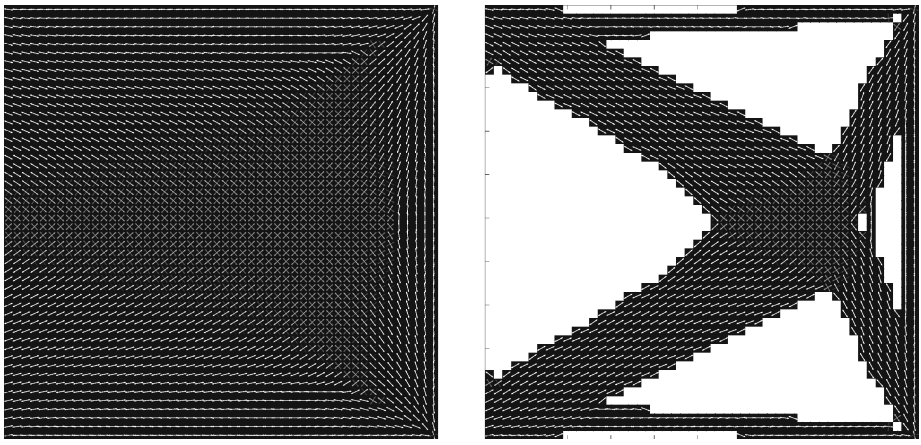


(a) $a = b = 0.02$ (b) square

Fig. 2. Optimization for $\mathcal{C}_{\theta,s}$ with similar results. The arrows denote the principal stiffness directions and the gray value the magnitude (black: low, white: high).

Table 2. Compliance values and FMO comparison for 50x50 squared FE patches.

$\mathcal{C}_{\theta,s=0}$	3.6850	38.89 %
$\mathcal{C}_{\theta,s}$	2.9963	12.93 %
$\mathcal{C}_{\text{Eng}}$	3.0535	15.08 %
$\mathcal{C}_{\text{FMO}}$	2.8964	9.16 %
FMO	2.6533	



(a) square (b) square, topology

Fig. 3. Results of the simultaneous material and topology optimization for $\mathcal{C}_{\theta,s}$.

subproblems $\mathcal{C}_{\text{Eng}}$, apparently a local minimum was found, which stresses the importance of a global solution of the local material optimization problems.

2.4 Material and Topology Optimization

The term (6) in the asymptotic expansion actually corresponds to a topological derivative and can hence be used for topology optimization. We propose an algorithm, where we do not only use the information of "drilling a hole", but use the values obtained in material optimization instead. The idea is, that those finite element patches, which provide the smallest gain, when substituting the optimal material C^0, are left out in the following iteration:

(S1) choose matrix material C^1 and admissible materials for inclusions $\mathcal{C}$;
(S2) compute Ψ from (4);
(S3) define loads and boundary conditions;
while *volume* $>$ *bound* **do**
  (S4) solve state problem without inclusions for current topology;
  foreach *inclusion* ω_i, $i = 1, \ldots, n_{ell}$ **do**
    (S5) solve $\min_{C_i^0 \in \mathcal{C}} \mathcal{D}_c(C_i^0, z_i)$ to global optimality;
  end
  (S6) remove FE-patches with $\min_{C_i^0 \in \mathcal{C}} \mathcal{D}_c(C_i^0, z_i)$ largest;
end

Algorithm 2. Basic algorithm for simultaneous material and topology optimization based on the asymptotic model.

Note that, again, in each iteration the state problem only needs to be solved once.

In Fig. 3(b), the result of the algorithm for the admissible material set $\mathcal{C}_{\theta,s}$ and 50×50 squared FE patches is shown. In this experiment, we removed in the iteration k a total of $200 * 0.83^k$ FE patches until 3 FE patches or less where removed, which lead to a final volume fraction of 0.542. The strategy for the removal of ellipses can be varied, however a decreasing volume fraction should be removed in order to obtain a smoother convergence.

3 Conclusion

We proposed an efficient algorithm for material optimization on multiple elliptic inclusions. The numerical evidence suggests that the accuracy of the proposed method decreases only slightly, when replacing all material instead of just the material in elliptic inclusions. A major advantage of this algorithm is the possibility of avoiding local minima, however at the cost of only having an approximate solution. The total error in the studied example for the finer resolution was about 10 %. From experience in practice, the proposed algorithm for simultaneous material and topology optimization seems to work well in large parts, however oftentimes small bars are left over and if a hole happens to be drilled in a "bad" position the algorithm struggles, as material is not reintroduced.

Nevertheless, both algorithms allow for a very efficient solution of usually quite complicated problems, such as discrete material optimization and rotational optimization. When the accuracy provided in this method does not suffice or the optimized topology appears flawed, the optimization result may still be used as a high quality initial design for other solution schemes, such as fully parametrized approaches.

References

1. Leugering, G., Nazarov, S., Schury, F., Stingl, M.: The Eshelby theorem and application to the optimization of an elastic patch. SIAM J. Appl. Math. **72**(2), 512–534 (2012)
2. Pedersen, P.: On optimal orientation of orthotropic materials. Struct. Multidiscip. Opt. **1**, 101–106 (1989)
3. Pedersen, P.: On thickness and orientational design with orthotropic materials. Struct. Multidiscip. Opt. **3**, 69–78 (1991)
4. Thomsen, J.: Optimization of composite discs. Struct. Multidiscip. Opt. **3**, 89–98 (1991)
5. Lewiński, T., Sokołowski, J.: Energy change due to the appearance of cavities in elastic solids. Int. J. Solids Struct. **40**(7), 1765–1803 (2003)
6. Schury, F., Greifenstein, J., Leugering, G., Stingl, M.: On the efficient solution of a patch problem with multiple elliptic inclusions. Optim. Eng. (Accepted for publication, 2014)
7. Love, A.E.H.: A Treatise on the Mathematical Theory of Elasticity. Dover Publications, New York (1944)
8. Haslinger, J., Kočvara, M., Leugering, G., Stingl, M.: Multidisciplinary free material optimization. SIAM J. Appl. Math. **70**(7), 2709–2728 (2010)
9. Stingl, M., Kočvara, M., Leugering, G.: A sequential convex semidefinite programming algorithm with an application to multiple-load free material optimization. SIAM J. Opt. **20**(1), 130–155 (2009)

Steady Fluid-Structure Interaction Using Fictitious Domain

Andrei Halanay[1(✉)] and Cornel Marius Murea[2]

[1] Department of Mathematics 1, University Politehnica of Bucharest,
313 Splaiul Independenţei, 060042 Bucharest, Romania
halanay@mathem.pub.ro
[2] Laboratoire de Mathématiques, Informatique et Applications
Université de Haute Alsace, 4-6 Rue des Fréres Lumière,
68093 Mulhouse Cedex, France
cornel.murea@uha.fr
http://www.edp.lmia.uha.fr/murea/

Abstract. We present a formulation for a steady fluid-structure interaction problem using fictitious domain technique with penalization. Numerical results are presented.

Keywords: Fluid-structure interaction · Fictitious domain

1 Setting for a Steady Fluid-Structure Interaction Problem

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 $\overline{\Omega}_0^S \subset D$.

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 $\overline{\Omega}_u^S \subset D$ and the fluid occupies $\Omega_u^F = D \backslash \overline{\Omega}_u^S$. We set $\Gamma_u = \varphi(\Gamma_0)$, then the boundary of the deformed structure is $\partial\Omega_u^S = \Gamma_D \cup \Gamma_u$ and the boundary of the fluid domain admits the decomposition $\partial\Omega_u^F = \partial D \cup \Gamma_D \cup \Gamma_u$. The fluid-structure geometrical configuration is represented in Fig. 1.

Generally, the fluid equations are described using Eulerian coordinates, while for the structure equations, the Lagrangian coordinates are employed. The gradients with respect to the Eulerian coordinates $\mathbf{x} \in \Omega_u^S$ of a scalar field $q : D \to \mathbb{R}$ or a vector field $\mathbf{w} = (w_1, w_2) : D \to \mathbb{R}^2$ are denoted by ∇q and $\nabla \mathbf{w}$. The divergence operators with respect to the Eulerian coordinates of a vector field

Andrei Halanay: Supported by Grant ID-PCE 2011-3-0211, Romania.

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 128–137, 2014.
DOI: 10.1007/978-3-662-45504-3_12

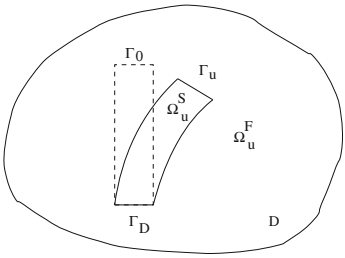


Fig. 1. Geometrical configuration.

$\mathbf{w} = (w_1, w_2) : D \to \mathbb{R}^2$ and of a tensor $\sigma = (\sigma_{ij})_{1 \le i,j \le 2}$ are denoted by $\nabla \cdot \mathbf{w}$ and $\nabla \cdot \sigma$.

When the derivatives are with respect to the Lagrangian coordinates $\mathbf{X} = \varphi^{-1}(\mathbf{x}) \in \Omega_0^S$, we use the notations: $\nabla_{\mathbf{X}} \mathbf{u}$, $\nabla_{\mathbf{X}} \cdot \mathbf{u}$, $\nabla_{\mathbf{X}} \cdot \sigma$.

If $\mathbf{A}$ is a square matrix, we denote by $\det \mathbf{A}$, $\mathbf{A}^{-1}$, $\mathbf{A}^T$ its determinant, the inverse and the transposed matrix, respectively. We write $\operatorname{cof} \mathbf{A} = (\det \mathbf{A}) \left(\mathbf{A}^{-1}\right)^T$ the co-factor matrix of $\mathbf{A}$. We write $\mathbf{A}^{-T} = \left(\mathbf{A}^{-1}\right)^T$.

We denote by $\mathbf{F}(\mathbf{X}) = \mathbf{I} + \nabla_{\mathbf{X}} \mathbf{u}(\mathbf{X})$ the gradient of the deformation and by $J(\mathbf{X}) = \det \mathbf{F}(\mathbf{X})$ the Jacobian determinant, where $\mathbf{I}$ is the unit matrix.

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.

The problem is to find the structure displacement $\mathbf{u} : \overline{\Omega}_0^S \to \mathbb{R}^2$, the fluid velocity $\mathbf{v} : \overline{\Omega}_u^F \to \mathbb{R}^2$ and the fluid pressure $p : \overline{\Omega}_u^F \to \mathbb{R}$ such that:

$$-\nabla_{\mathbf{X}} \cdot \sigma^S(\mathbf{u}) = \mathbf{f}^S, \quad \text{in } \Omega_0^S \tag{1}$$

$$\mathbf{u} = 0, \quad \text{on } \Gamma_D \tag{2}$$

$$-\nabla \cdot \sigma^F(\mathbf{v}, p) = \mathbf{f}^F, \quad \text{in } \Omega_u^F \tag{3}$$

$$\nabla \cdot \mathbf{v} = 0, \quad \text{in } \Omega_u^F \tag{4}$$

$$\mathbf{v} = 0, \quad \text{on } \partial D \tag{5}$$

$$\mathbf{v} = 0, \quad \text{on } \Gamma_D \tag{6}$$

$$\mathbf{v} = 0, \quad \text{on } \Gamma_u \tag{7}$$

$$\omega\left(\sigma^F(\mathbf{v}, p)\,\mathbf{n}^F\right) \circ \varphi = -\sigma^S(\mathbf{u})\,\mathbf{n}^S, \quad \text{on } \Gamma_0 \tag{8}$$

where $\mathbf{f}^S : \Omega_0^S \to \mathbb{R}^2$ are the applied volume forces on the structure and $\mathbf{n}^S$ is the structure unit outward vector normal to $\partial\Omega_0^S$. Similarly, we define $\mathbf{f}^F : \Omega_u^F \to \mathbb{R}^2$ and $\mathbf{n}^F$ the fluid unit outward vector normal to $\partial\Omega_u^S$.

We point out that the stress tensor of the structure is defined on the undeformed structure domain Ω_0^S, while the Cauchy stress tensors of the fluid is defined in the deformed domain Ω_u^F.

We have used the notation $\omega(\mathbf{X}) = \left\|J\,\mathbf{F}^{-T}\mathbf{n}^S\right\|_{\mathbb{R}^2} = \left\|\mathrm{cof}\,(\mathbf{F})\,\mathbf{n}^S\right\|_{\mathbb{R}^2}$ for $\mathbf{X}$ on $\partial\Omega_0^S$, which is a kind of Jacobian determinant for the change of variable formula for integral over surface.

The Eqs. (1), (2) concern the structure, while (3)–(6) concern the fluid. The Eqs. (7), (8) represent the boundary conditions on the moving fluid-structure interface. The fluid and the structure domains Ω_u^F, Ω_u^S depend on the structure displacement u which is unknown.

2 Parametrization and Regularization of the Characteristic Function

The regularization of the characteristic function of the deformed structure domain is necessary in order to prove the continuity of the solution with respect to the structure displacement, [3].

Denote by $\|\cdot\|_{1,\infty,\Omega}$ the usual norm of the Sobolev space $W^{1,\infty}(\Omega)$ and by $\|\cdot\|_{m,\Omega}$ the usual norm of $H^m(\Omega)$, $m \geq 0$ with the convention $H^0(\Omega) = L^2(\Omega)$.

For every $0 < \delta < 1$, there exists $0 < \eta_\delta < 1$ such that

$$1 - \delta \leq \det(\mathbf{I} + \nabla\mathbf{u}) \leq 1 + \delta, \quad \text{a.e. } \mathbf{x} \in \Omega_0^S \tag{9}$$

for all $\mathbf{u} \in \left(W^{1,\infty}(\Omega_0^S)\right)^2$ that satisfy $\|\mathbf{u}\|_{1,\infty,\Omega_0^S} \leq \eta_\delta$.

We define

$$B_\delta = \{u \in W^{1,\infty}(\Omega_0^S)^2; ||u||_{1,\infty,\Omega_0^S} \leq \eta_\delta,\ u = 0 \text{ on } \Gamma_D\}. \tag{10}$$

Let $j \in W^{1,\infty}(D)$ be a parametrization of $\Omega_0^S \subset D$, i.e. :

$$j(\mathbf{x}) > 0,\ \mathbf{x} \in \Omega_0^S, \quad j(\mathbf{x}) < 0,\ \mathbf{x} \in D\backslash\overline{\Omega}_0^S, \quad j(\mathbf{x}) = 0,\ \mathbf{x} \in \partial\Omega_0^S.$$

The parametrization is not necessarily unique.

Let $\mathbf{u} \in B_\delta$ be a given structure displacement. Denote, as before, $\Omega_u^S = \varphi(\Omega_0^S)$, where $\varphi(\mathbf{X}) = \mathbf{X} + \mathbf{u}(\mathbf{X})$. Then $\varphi : \overline{\Omega}_0^S \to \overline{\Omega}_u^S$ is bijective and bilipschitzian and

$$j_u(\mathbf{y}) = \begin{cases} j(\mathbf{x}), & \mathbf{y} = \varphi(\mathbf{x}) \in \Omega_u^S \\ 0, & \mathbf{y} \in \partial\Omega_u^S \\ -dist(\mathbf{y}, \overline{\Omega}_u^S), & \mathbf{y} \notin \overline{\Omega}_u^S \end{cases}$$

is a parametrization of Ω_u^S, $j_u \in W^{1,\infty}(D)$.

If H is the Heaviside function $H : \mathbb{R} \to \{0, 1\}$,

$$H(r) = \begin{cases} 1, r \geq 0 \\ 0, r < 0 \end{cases}$$

then $H(j_u(\cdot))$ is the characteristic function of Ω_u^S.

For $\varepsilon > 0$, let $\Omega_0^\varepsilon \subset\subset \Omega_0^S$. Since $j : D \to \mathbb{R}$ is Lipschitz continuous and $j > 0$ in Ω_0^S, there is $\mu_\varepsilon > 0$ such that $j(\mathbf{x}) \geq \mu_\varepsilon > 0$, for all $\mathbf{x} \in \Omega_0^\varepsilon$. Consequently,

$$\mu_\varepsilon \leq \min_{\mathbf{y} \in \overline{\Omega_u^\varepsilon}} j_u(y), \quad \forall \mathbf{u} \in B_\delta.$$

Then we take H^{μ_ε}, the Yosida regularization of H

$$H^{\mu_\varepsilon}(r) = \begin{cases} 1, & r \geq \mu_\varepsilon \\ \frac{r}{\mu_\varepsilon} & 0 \leq r < \mu_\varepsilon \\ 0, & r < 0 \end{cases}$$

and we set $\widetilde{H}_u(\mathbf{x}) = H^{\mu_\varepsilon}(j_u(\mathbf{x}))$ for all $\mathbf{x} \in D$, which is a Lipschitz regularization of the characteristic function of Ω_u^S. We have constructed $\widetilde{H}_u : \overline{D} \to \mathbb{R}$, Lipschitz on $\overline{D}$, $0 \leq \widetilde{H}_u(\mathbf{x}) \leq 1$, for all $\mathbf{x}$ in $\overline{D}$ such that

$$\widetilde{H}_u(\mathbf{x}) = \begin{cases} 0, \mathbf{x} \in \overline{D} \backslash \Omega_u^S \\ 1, \mathbf{x} \in \Omega_u^\varepsilon \end{cases} \tag{11}$$

where $\Omega_u^\varepsilon \subset\subset \Omega_u^S$.

3 Weak Formulation Using Fictitious Domain Technique with Penalization

We assume that D and Ω_0^S are Lipschitz. Let us introduce the bi-linear forms

$$a_S\left(\mathbf{u}, \mathbf{w}^S\right) = \int_{\Omega_0^S} \left(\lambda^S \left(\nabla \cdot \mathbf{u}\right)\left(\nabla \cdot \mathbf{w}^S\right) + 2\mu^S \epsilon\left(\mathbf{u}\right) : \epsilon\left(\mathbf{w}^S\right)\right) d\mathbf{X}$$

$$a_F\left(\mathbf{v}, \mathbf{w}\right) = \int_D 2\mu^F \epsilon\left(\mathbf{v}\right) : \epsilon\left(\mathbf{w}\right) d\mathbf{x}$$

$$b_F\left(\mathbf{w}, p\right) = -\int_D \left(\nabla \cdot \mathbf{w}\right) p \, d\mathbf{x}$$

and the Hilbert spaces

$$W^S = \left\{\mathbf{w}^S \in \left(H^1\left(\Omega_0^S\right)\right)^2;\ \mathbf{w}^S = 0 \text{ on } \Gamma_D\right\},$$

$$W = \left(H_0^1\left(D\right)\right)^2,$$

$$Q = L_0^2\left(D\right) = \{q \in L^2\left(D\right);\ \int_D q \, dx = 0\}.$$

We assume that $\mathbf{f}^F \in \left(L^2(D)\right)^2$, $\mathbf{f}^S \in \left(L^2(\Omega_0^S)\right)^2$.

Weak fluid formulation using fictitious domain. For a given $\mathbf{u} \in B_\delta$, we define: fluid velocity $\mathbf{v}_\varepsilon \in W$ and fluid pressure $p_\varepsilon \in Q$, as the solution of the following problem:

$$a_F\left(\mathbf{v}_\varepsilon, \mathbf{w}\right) + b_F\left(\mathbf{w}, p_\varepsilon\right)$$

$$+\frac{1}{\varepsilon}\int_D \widetilde{H}_u\left(\mathbf{v}_\varepsilon \cdot \mathbf{w} + \nabla \mathbf{v}_\varepsilon : \nabla \mathbf{w}\right) d\mathbf{x} = \int_D \mathbf{f}^F \cdot \mathbf{w} d\mathbf{x}, \forall \mathbf{w} \in W \tag{12}$$

$$b_F\left(\mathbf{v}_\varepsilon, q\right) = 0, \forall q \in Q \tag{13}$$

Weak structure formulation. For given $\mathbf{u} \in B_\delta$, $\mathbf{v}_\varepsilon \in W$ and $p_\varepsilon \in Q$, we define the structure displacement $\mathbf{u}_\varepsilon \in W^S$ as the solution of

$$a_S\left(\mathbf{u}_\varepsilon, \mathbf{w}^S\right) = \int_{\Omega_0^S} \mathbf{f}^S \cdot \mathbf{w}^S \, d\mathbf{X} + \int_{\Omega_0^S} J\left(\sigma^F\left(\mathbf{v}_\varepsilon, p_\varepsilon\right) \circ \varphi\right) \mathbf{F}^{-T} : \nabla_{\mathbf{X}} \mathbf{w}^S \, d\mathbf{X}$$
$$+ \frac{1}{\varepsilon} \int_{\Omega_0^S} J \widetilde{H}_u(\varphi) \left(\left(\mathbf{v}_\varepsilon \circ \varphi\right) \cdot \mathbf{w}^S + \left(\nabla \mathbf{v}_\varepsilon \circ \varphi\right) \mathbf{F}^{-T} : \nabla_{\mathbf{X}} \mathbf{w}^S \right) d\mathbf{X}$$
$$- \int_{\Omega_0^S} J\left(\mathbf{f}^F \circ \varphi\right) \cdot \mathbf{w}^S \, d\mathbf{X}, \ \forall \mathbf{w}^S \in W^S \quad (14)$$

where $\varphi(\mathbf{X}) = \mathbf{X} + \mathbf{u}(\mathbf{X})$, $\mathbf{F}(\mathbf{X}) = \mathbf{I} + \nabla_{\mathbf{X}} \mathbf{u}(\mathbf{X})$, $J(\mathbf{X}) = \det \mathbf{F}(\mathbf{X})$.

Remark 1. From the structure equation $-\nabla \cdot \sigma^S\left(\mathbf{u}_\varepsilon\right) = \mathbf{f}^S$, in Ω_0^S using Green's formula, we obtain for all $\mathbf{w}^S = 0$ on Γ_D that

$$a_S\left(\mathbf{u}_\varepsilon, \mathbf{w}^S\right) = \int_{\Omega_0^S} \mathbf{f}^S \cdot \mathbf{w}^S \, d\mathbf{X} + \int_{\Gamma_0} \sigma^S\left(\mathbf{u}_\varepsilon\right) \mathbf{n}^S \cdot \mathbf{w}^S dS.$$

We can prove (see [3]) that the sum of the last three terms in (14) is equal to the fluid forces acting on the structure which is also equals to $\int_{\Gamma_0} \sigma^S\left(\mathbf{u}_\varepsilon\right) \mathbf{n}^S \cdot \mathbf{w}^S dS$. In fact, from (14) and the above weak formulation of the structure, we can get that the boundary condition at the interface concerning the continuity of the stress (8) is verified in a weak sense (see [3]). The second boundary condition at the interface is the continuity of the velocity (7). This is obtained by using the penalization term in the structure domain in (12).

For each $i \in \mathbb{N}^*$, there exists an unique eigenvalue $\lambda_i > 0$ and an unique eigenfunction $\boldsymbol{\phi}^i \in \mathbf{W}^S$, solution of

$$a_S\left(\boldsymbol{\phi}^i, \mathbf{w}^S\right) = \lambda_i \int_{\Omega_0^S} \boldsymbol{\phi}^i \cdot \mathbf{w}^S d\mathbf{X}, \quad \forall \mathbf{w}^S \in \mathbf{W}^S \quad (15)$$

such that

$$\int_{\Omega_0^S} \boldsymbol{\phi}^i \cdot \boldsymbol{\phi}^j d\mathbf{X} = \delta_{ij}, \quad (16)$$

see [8], Chap. [6]. We assume that $\lambda_1 \leq \lambda_2 \leq \ldots$. The set $\{\boldsymbol{\phi}^i, \ i \in \mathbb{N}^*\}$ forms an orthonormal basis of $L^2(\Omega_0^S)$. Let $m \in \mathbb{N}^*$ be given. Let $\mathbf{u}_\varepsilon^m$ be the orthogonal projection of $\mathbf{u}_\varepsilon$ on $span\left(\boldsymbol{\phi}^i, \ i = 1, \ldots, m\right)$ in $L^2(\Omega_0^S)$, so $\mathbf{u}_\varepsilon^m = \sum_{i=1}^m \alpha_i \boldsymbol{\phi}^i$, $\alpha_i \in \mathbb{R}$.

We define

$$B_\delta^m = \Big\{ \mathbf{u} \in \left(W^{1,\infty}(\Omega_0^S)\right)^2; \ \mathbf{u} = 0 \text{ on } \Gamma_D, \ \|\mathbf{u}\|_{1,\infty,\Omega_0^S} < \eta_\delta,$$
$$\mathbf{u} = \sum_{i=1}^m \alpha_i \boldsymbol{\phi}^i, \ \alpha_i \in \mathbb{R} \Big\} . \quad (17)$$

We have $B_\delta^m \subset B_\delta$. For each $\mathbf{u} \in B_\delta$, we define the nonlinear operator

$$T_\varepsilon^m(\mathbf{u}) = \mathbf{u}_\varepsilon^m.$$

A solution of the penalized fluid-structure interaction problem will be, by definition, a fixed point of T_ε^m in B_δ^m.

Remark 2. As in [3], we can prove the existence of a solution of the penalized fluid-structure interaction problem $T_\varepsilon^m(\mathbf{u}_\varepsilon^m) = \mathbf{u}_\varepsilon^m$ using the Schauder fixed point theorem. In order to obtain supplementary regularity of the Stokes equations, a non linear penalization term $\frac{1}{\varepsilon}\int_D \widetilde{H}_u \, sgn(\mathbf{v}_\varepsilon)\,|\mathbf{v}_\varepsilon|^{\alpha-1} \cdot \mathbf{w}\, d\mathbf{x}$ was employed in [3], where $\alpha > 2$. In the present paper, we use a linear penalization term in (12), but we are working only with a finite number of eigenfunctions of the structure equations. The behavior of $\mathbf{u}_\varepsilon^m$ when ε goes to zero will be studied in [4].

4 Fixed Point Iterations

We replace $\widetilde{H}_u$ in (12) and (14) by χ_u^S the characteristic function of Ω_u^S in order to simplify the computation. The regularization of the characteristic function was necessary to obtain 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 unit matrix $\mathbf{I}$.

Algorithm 1 by fixed point iterations

Step 1. Given the initial displacement of the structure $\mathbf{u}^{m,0} = \sum_{i=1}^m \alpha_i^0 \boldsymbol{\phi}^i$, compute the characteristic function $\chi_{u^0}^S$, put $k := 0$.

Step 2. Find the velocity $\mathbf{v}_\varepsilon^k \in \left(H^1(D)\right)^2$, $\mathbf{v}_\varepsilon^k = \mathbf{g}$ on ∂D and the pressure $p_\varepsilon^k \in Q$ by solving the fluid problem

$$\begin{aligned} a_F\left(\mathbf{v}_\varepsilon^k, \mathbf{w}\right) + b_F\left(\mathbf{w}, p_\varepsilon^k\right) \\ + \frac{1}{\varepsilon}\int_D \chi_{u_\varepsilon^k}^S \left(\mathbf{v}_\varepsilon^k \cdot \mathbf{w} + \nabla \mathbf{v}_\varepsilon^k : \nabla \mathbf{w}\right) d\mathbf{x} = \int_D \mathbf{f}^F \cdot \mathbf{w}\, d\mathbf{x}, \ \forall \mathbf{w} \in W \\ b_F\left(\mathbf{v}_\varepsilon^k, q\right) = 0, \ \forall q \in Q. \end{aligned}$$

Step 3. Find the new displacement of the structure $\mathbf{u}_\varepsilon^{m,k+1} = \sum_{i=1}^m \alpha_i^{k+1} \boldsymbol{\phi}^i$ by solving

$$\begin{aligned} \alpha_i^{k+1}\lambda_i = & \int_{\Omega_0^S} \left(\mathbf{f}^S - \mathbf{f}^F\right) \cdot \boldsymbol{\phi}^i \, d\mathbf{x} \\ & + \int_{\Omega_0^S} 2\mu^F \epsilon\left(\mathbf{v}_\varepsilon^k\right) : \epsilon\left(\boldsymbol{\phi}^i\right) \, d\mathbf{x} - \int_{\Omega_0^S} \left(\nabla \cdot \boldsymbol{\phi}^i\right) p_\varepsilon^k \, d\mathbf{x} \\ & + \frac{1}{\varepsilon}\int_{\Omega_0^S} \left(\left(\mathbf{v}_\varepsilon^k \circ \varphi_\varepsilon^k\right) \cdot \boldsymbol{\phi}^i + \left(\nabla \mathbf{v}_\varepsilon^k \circ \varphi_\varepsilon^k\right) : \nabla \boldsymbol{\phi}^i\right) d\mathbf{x}, \ i = 1, \ldots, m \end{aligned}$$

where $\varphi_\varepsilon^k(\mathbf{X}) = \mathbf{X} + \mathbf{u}_\varepsilon^{m,k}(\mathbf{X})$.

Step 4. Stopping test: if $\left\|\mathbf{u}_\varepsilon^{m,k}-\mathbf{u}_\varepsilon^{m,k+1}\right\|_{0,\Omega_0^S}\leq tol$, then **Stop**.
Step 5. Compute the characteristic function $\chi^S_{u_\varepsilon^{m,k+1}}$, put $k:=k+1$ and **Go to Step 2.**

A similar fixed point algorithm was used in [7].

5 Least Squares Approach

The previous algorithm converges if the operator T_ε^m is a contraction. But, the **Algorithm 1** fails for some physical parameters. For this reason, we introduce a second algorithm which is more robust.

For $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_m)\in\mathbb{R}^m$, we can define $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_m)\in\mathbb{R}^m$ by

$$T_\varepsilon^m\left(\sum_{i=1}^m\alpha_i\boldsymbol{\phi}^i\right)=\sum_{i=1}^m\beta_i\boldsymbol{\phi}^i.$$

We set the cost function $J(\boldsymbol{\alpha})=\frac{1}{2}\sum_{i=1}^m(\alpha_i-\beta_i)^2$ and now the problem to be solved is $\inf_{\boldsymbol{\alpha}\in\mathbb{R}^m}J(\boldsymbol{\alpha})$. In order to solve the optimization problem, we employ the quasi-Newton iterative method called Broyden, Fletcher, Goldforb, Shano (BFGS) scheme (see for example [2], Chap. [9]).

Algorithm 2 by the BFGS method
Step 0. Choose a starting point $\boldsymbol{\alpha}^0\in\mathbb{R}^m$, an $m\times m$ symmetric positive matrix H_0. Set $k=0$.
Step 1. Compute $\nabla J(\boldsymbol{\alpha}^k)$.
Step 2. If $\left\|\nabla J(\boldsymbol{\alpha}^k)\right\|<tol$ **Stop**.
Step 3. Set $\mathbf{d}^k=-H_k\nabla J(\boldsymbol{\alpha}^k)$.
Step 4. Determine $\boldsymbol{\alpha}^{k+1}=\boldsymbol{\alpha}^k+\theta_k\mathbf{d}^k$, $\theta_k>0$ by means of an approximate minimization

$$J(\boldsymbol{\alpha}^{k+1})\approx\min_{\theta\geq0}J(\boldsymbol{\alpha}^k+\theta\mathbf{d}^k).$$

Step 5. Compute $\delta_k=\boldsymbol{\alpha}^{k+1}-\boldsymbol{\alpha}^k$.
Step 6. Compute $\nabla J(\boldsymbol{\alpha}^{k+1})$ and $\gamma_k=\nabla J(\boldsymbol{\alpha}^{k+1})-\nabla J(\boldsymbol{\alpha}^k)$.
Step 7. Compute

$$H_{k+1}=H_k+\left(1+\frac{\gamma_k^TH_k\gamma_k}{\delta_k^T\gamma_k}\right)\frac{\delta_k\delta_k^T}{\delta_k^T\gamma_k}-\frac{\delta_k\gamma_k^TH_k+H_k\gamma_k\delta_k^T}{\delta_k^T\gamma_k}$$

Step 8. Update $k=k+1$ and go to the **Step 2**.

The matrices H_k approach the inverse of the Hessian of J. For the inaccurate line search at the **Step 4**, the methods of Goldstein and Armijo were used. If we denote by $g:[0,\infty)\to\mathbb{R}$ the function $g(\theta)=J(\boldsymbol{\alpha}^k+\theta\mathbf{d}^k)$, we determine $\theta_k>0$ such that: $g(0)+(1-\lambda)\,\theta_kg'(0)\leq g(\theta_k)\leq g(0)+\lambda\theta_kg'(0)$, where $\lambda\in(0,1/2)$.

In this paper, we compute $\nabla J(\boldsymbol{\alpha})$ by the Finite Differences Method $\frac{\partial J}{\partial\alpha_k}(\boldsymbol{\alpha})\approx\left(J(\boldsymbol{\alpha}+\Delta\alpha_k\mathbf{e_k})-J(\boldsymbol{\alpha})\right)/\Delta\alpha_k$, where $\mathbf{e_k}$ is the k-th vector of the canonical base of $\mathbb{R}^m$ and $\Delta\alpha_k>0$ is the grid spacing.

Concerning the convergence rate, the fixed point algorithm is slower than the BFGS Method. If the starting point is not sufficiently close to the solution, the fixed point algorithm diverges. On the contrary, the BFGS Method is less sensitive to the choice of the starting point and, in general, it is convergent to a local minimizer from almost any starting point. This is the main advantage, (see [6]).

6 Numerical Results. Deformation of a Tall Building Under the Action of Wind

We have performed numerical simulations using a 2D model adapted from [1] (see Fig. 2).

The dimensions of a rectangular tall building are: height $H = 180\,\mathrm{m}$, length $L = 30\,\mathrm{m}$. The computational domain of the fluid D is a rectangle of height $H_1 = 3H$ and length $L_1 = L + 4H$, its left bottom corner is at $(0, 0)$. We shall allow nonhomogeneous Dirichlet data in the numerical experiments.

The distance between the left side of the fluid and the left side of the structure is H. We denote by Σ_1, Σ_3 the left and the right vertical boundaries and by Σ_2, Σ_4 the bottom and the top boundaries, respectively.

The mechanical properties of the building assumed to be an elastic structure are: Young modulus $E^S = 2.3 \times 10^5\,\mathrm{N/m^2}$, Poisson's ratio $\nu^S = 0.25$, the applied volume forces on the structure $\mathbf{f}^S : \Omega_0^S \to \mathbb{R}^2$, $\mathbf{f}^S = (0, 0)\,\mathrm{N/m^3}$. If the Young modulus is $E^S = 2.3 \times 10^8\,\mathrm{N/m^2}$ as in [1], the displacements of the structure are very small.

The fluid is the air with: dynamic viscosity $\mu^F = 7.03 \times 10^{-2}\,\mathrm{N \cdot s/m^2}$, the applied volume forces on the fluid $\mathbf{f}^S : D \to \mathbb{R}^2$, $\mathbf{f}^F = (0, 0)\ \mathrm{N/m^3}$. The inflow velocity profile is $\mathbf{g}(x_1, x_2) = 100\left(\frac{x_2}{H}\right)^{0.19}\ \mathrm{m/s}$. The considered boundary conditions for the fluid are more natural from the point of view of applications and differ slightly compared with the previous sections. We impose: $\mathbf{v}_\epsilon = \mathbf{g}$ on $\Sigma_1 \cup \Sigma_4$, $\mathbf{v}_\epsilon = 0$ on Σ_2 and $\sigma^F(\mathbf{v}, p)\,\mathbf{n}^F = 0$ on Σ_3.

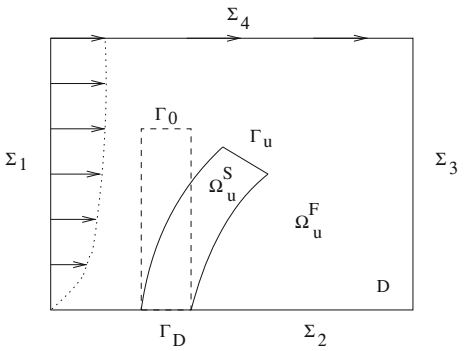


Fig. 2. Geometrical configuration for the numerical results

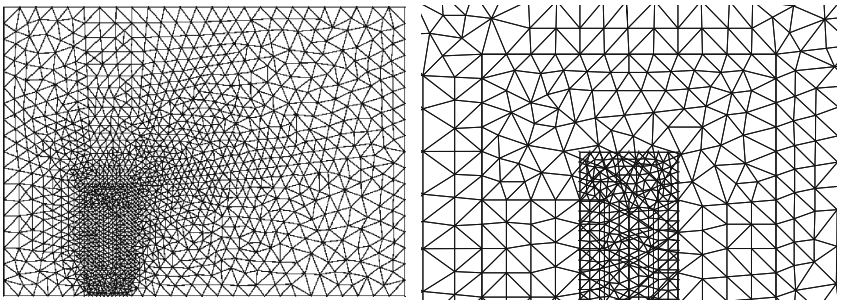


Fig. 3. The fixed mesh of the fluid domain (left). The finer zone $[H-L, H+2L] \times [0, H+L]$ of the fluid mesh covers the structure mesh which occupies initially the rectangle $[H, H+L] \times [0, H]$ (zoom, right). The fluid and structure meshes are not compatible, for example, a vertex on the structure boundary is not necessary a vertex on the fluid mesh (right).

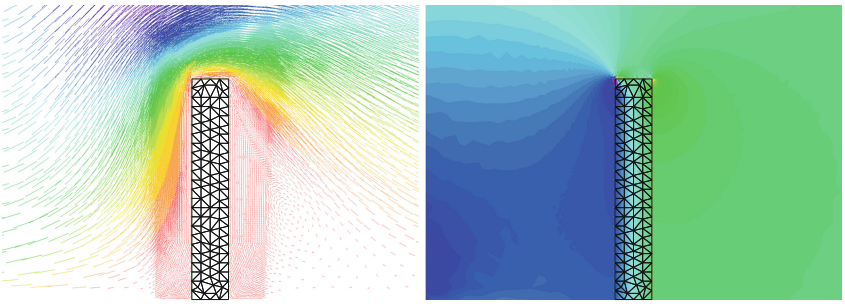


Fig. 4. Velocity (left) and pressure (right) of the fluid around the deformed structure.

The numerical tests have been produced using the software *FreeFem++* [5]. 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 on a mesh of 34871 triangles and 17550 vertices. The finite element $\mathbb{P}_1$ was used in order to solve the structure problem on a mesh of 192 triangles and 125 vertices. The characteristic function was approached by $\mathbb{P}_0$ finite element.

We have performed the simulation using the **Algorithm 2** described in the previous section. We have used the initial displacement $\alpha^0 = 0$ at the **Step 0** and the tolerance $tol = 0.0001$ for the stopping criterion at the **Step 2**. The penalization parameter is $\varepsilon = 0.001$ and the number of the eigenfunctions is $m = 5$. The stopping criterion holds after 6 iterations of the BFGS algorithm, the initial value of the cost function is 34.30 and the final value is 5.9×10^{-13}. The maximal structural displacement is 0.148 m.

The fluid velocity is almost zero in the deformed structure domain, more precisely $\|\mathbf{v}_\varepsilon\|_{1,\Omega^S_{u_\varepsilon}} = \sqrt{\int_D \chi^S_{u_\varepsilon} \left(\mathbf{v}_\varepsilon \cdot \mathbf{v}_\epsilon + \nabla \mathbf{v}_\varepsilon : \nabla \mathbf{v}_\varepsilon\right) d\mathbf{x}} = 0.00555$.

References

1. Braun, A.L., Awruch, A.M.: Aerodynamic and aeroelastic analyses on the CAARC standard tall building model using numerical simulation. Comput. Struct. **87**, 564–581 (2009)
2. Dennis, Jr., J.E., Schnabel, R.B.: Numerical methods for unconstrained optimization and nonlinear equations. Classics in Applied Mathematics, vol. 16 (1996) (Society for Industrial and Applied Mathematics, Philadelphia, PA)
3. Halanay, A., Murea, C.M., Tiba, D.: Existence and approximation for a steady fluid-structure interaction problem using fictitious domain approach with penalization. Math. Appl. **5**(1–2), 120–147 (2013)
4. Halanay, A., Murea, C.M.: Existence of a steady flow of Stokes fluid past a linear elastic structure using fictitious domain (in preparation)
5. Hecht, F.: http://www.freefem.org
6. Murea, C.M.: Numerical simulation of a pulsatile flow through a flexible channel. ESAIM: Math. Model. Numer. Anal. **40**(6), 1101–1125 (2006)
7. Murea, C.M., Halanay, A.: Embedding domain technique for a fluid-structure interaction problem. In: Hömberg, D., Tröltzsch, F. (eds.) System Modeling and Optimization. IFIP AICT, vol. 391, pp. 358–367. Springer, Heidelberg (2013)
8. Raviart, P.-A., Thomas, J.-M.: Introduction à l'analyse numérique des équations aux dérivées partielles. Dunod, Paris (1998)

Sensitivity of the Solution Set to Second Order Evolution Inclusions

Jiangfeng Han and Stanislaw Migorski(✉)

Faculty of Mathematics and Computer Science, Institute of Computer Science, Jagiellonian University, ul. Łojasiewicza 6, 30348 Krakow, Poland
migorski@ii.uj.edu.pl

Abstract. In this note we study second order evolution inclusions in the framework of evolution triple of spaces. The existence of mild solutions (i.e. trajectory-selection pairs) to the inclusion, and the upper and lower semicontinuity properties of the solution set with respect to a parameter are established.

Keywords: Evolution inclusion · Kuratowski convergence · Upper semicontinuity · Lower semicontinuity

1 Introduction and Preliminaries

In this paper we investigate a class of systems described by abstract second order evolution equations with multivalued right hand side. We consider Problem (P) of the form

$$\begin{cases} \ddot{x}(t) + A(t, \dot{x}(t)) + Bx(t) \in F(t, x(t), \dot{x}(t)) \ \text{ a.e. } t \in (0,T), \\ x(0) = x_0, \ \ \dot{x}(0) = x_1 \end{cases}$$

and the following sequence of Problems $(\mathrm{P})_n$, $n \in \mathbb{N}$, that can be regarded as the perturbed ones

$$\begin{cases} \ddot{x}(t) + A_n(t, \dot{x}(t)) + B_n x(t) \in F_n(t, x(t), \dot{x}(t)) \ \text{ a.e. } t \in (0,T), \\ x(0) = x_0^n, \ \ \dot{x}(0) = x_1^n \,. \end{cases}$$

The goal is to establish the lower and upper semicontinuity properties of the solution set to Problem (P) with respect to the parameter $n \in \mathbb{N}$. The main result concerns the Kuratowski convergence of the sequence of solution sets to

This research was supported by the Marie Curie International Research Staff Exchange Scheme Fellowship within the 7th European Community Framework Programme under Grant Agreement No. 295118, the National Science Center of Poland under grant no. N N201 604640, the International Project co-financed by the Ministry of Science and Higher Education of Republic of Poland under grant no. W111/7.PR/2012, the National Science Center of Poland under Maestro Advanced Project no. DEC-2012/06/A/ST1/00262.

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 138–147, 2014.
DOI: 10.1007/978-3-662-45504-3_13

Problem $(P)_n$ to that of Problem (P). Evolution inclusions of second order and their applications have been considered in several papers, see e.g. [6–9] and the references therein.

We introduce below the notation and preliminary material needed in the next sections. For a Banach space X, we indicate by $w{-}X$, $s{-}X$ the space X equipped with the weak and the strong (norm) topology, respectively. Let (Ω, Σ, μ) be a measure space. A multifunction F defined on Ω with values in the space 2^X of all nonempty subsets of X is called measurable if $F^-(E) = \{\omega \in \Omega \mid F(\omega) \cap E \neq \emptyset\} \in \Sigma$ for every closed set $E \subset X$. It is called graph measurable if $GrF = \{(\omega, x) \in \Omega \times X \mid x \in F(\omega)\} \in \Sigma \times \mathcal{B}(X)$ where $\mathcal{B}(X)$ is the family of all Borel subsets of X. We denote by S_F^r, $1 \leq r \leq \infty$, the set of all selectors of F that belong to $L^r(\Omega; X)$, i.e., $S_F^r = \{f \in L^r(\Omega; X) \mid f(\omega) \in F(\omega)\ \mu \text{ a.e.}\}$. The symbol $\mathcal{P}_{f(c)}(X)$ stands for the family of all closed, (convex) subsets of 2^X. On $\mathcal{P}_f(X)$ we define the Hausdorff metric, by setting $h(A, B) = \max\{\sup_{a\in A} d(a, B), \sup_{b\in B} d(b, A)\}$. We also write $|A| = \sup\{|a| \mid a \in A\}$.

Given $\{S_n, S\}_{n\in\mathbb{N}} \subset 2^Z$, we define (see e.g. [3]) the sequential Kuratowski lower and upper limits respectively by $\tau_Z{-}\liminf S_n = \{z \in Z \mid \exists\, z_n \in S_n, z_n \to z \text{ in } \tau_Z{-}Z, \text{ as } n \to +\infty\}$ and $\tau_Z{-}\limsup S_n = \{z \in Z \mid \exists\, \{n_\nu\}, z_{n_\nu} \in S_{n_\nu}, z_{n_\nu} \to z \text{ in } \tau_Z{-}Z, \text{ as } \nu \to +\infty\}$. We say that S_n converge to S in the Kuratowski sense (denoted by $S_n \xrightarrow{K} S$) if and only if $\tau_Z{-}\limsup S_n \subset S \subset \tau_Z{-}\liminf S_n$.

Let (Y, τ_Y) and (Z, τ_Z) be Hausdorff topological spaces. A multifunction $G\colon Y \to 2^Z$ is said to be $(\tau_Y{-}\tau_Z)$ upper semicontinuous (usc) (respectively lower semicontinuous (lsc)) (cf. [2], Sect. 4.7 of [3]), if for every $C \subset Z$ closed in τ_Z topology, $G^-(C)$ (respectively, $G^+(C) = \{y \in Y \mid G(y) \subset C\}$) is closed in τ_Y topology in Y. The definition of lsc is equivalent to saying that if $y_n \to y$ in $\tau_Y{-}Y$, then $G(y) \subset \tau_Z{-}\liminf G(y_n)$. For a sequence of multifunctions G, $G_n\colon Y \to 2^Z$, we write

$$K(\tau_Y, \tau_Z) \limsup_{n\to+\infty, y\to\tilde{y}} G_n(y) \subset G(\tilde{y})$$

if $\tau_Z{-}\limsup G_n(y_n) \subset G(y)$ for every $y_n \to y$ in $\tau_Y{-}Y$. Similar notation is used for $\tau_Z{-}\liminf$.

Let H be a separable Hilbert space and V be a reflexive Banach space which is densely, continuously and compactly embedded in H. Identifying H with its dual H^*, we have the Gelfand triple $V \subset H \subset V^*$, where V^* is the dual of V. Let $\langle\cdot,\cdot\rangle$ be the duality of V and V^* as well as the inner product on H, let $\|\cdot\|$, $|\cdot|$ and $\|\cdot\|_{V^*}$ denote the norms in V, H and V^*, respectively. For $T > 0$ and $2 \leq p < +\infty$, we introduce the following spaces $\mathcal{V} = L^p(0, T; V)$, $\mathcal{H} = L^p(0, T; H)$, $\mathcal{H}^* = L^q(0, T; H)$, $\mathcal{V}^* = L^q(0, T; V^*)$, where $1/p + 1/q = 1$, $1 < q \leq 2$, and $\mathcal{W} = \{w \in \mathcal{V} \mid w' \in \mathcal{V}^*\}$. The derivative is understood in the sense of vector valued distributions. Clearly $\mathcal{W} \subset \mathcal{V} \subset \mathcal{H} \subset \mathcal{H}^* \subset \mathcal{V}^*$. The pairing of $\mathcal{V}$ and $\mathcal{V}^*$ and the duality between $\mathcal{H}$ and $\mathcal{H}^*$ are denoted by $\langle\langle f, v\rangle\rangle = \int_0^T \langle f(s), v(s)\rangle\, ds$. It is well known that the embedding $\mathcal{W} \subset C(0, T; H)$ is continuous. Since $V \subset H$ compactly we know that the embedding $\mathcal{W} \subset \mathcal{H}$ is

also compact. Finally, the class of linear bounded operators from V into V^* is denoted by $\mathcal{L}(V, V^*)$. For additional details on the material, we refer to [3,11].

2 Results on Evolution Equations

In this section we investigate the existence, uniqueness and continuous dependence of solutions on the data for an evolution equation of second order. We consider the following problem

$$\text{(E)} \qquad \begin{cases} \ddot{x}(t) + A(t, \dot{x}(t)) + Bx(t) = f(t) \ \text{ a.e. } t \in (0,T), \\ x(0) = x_0, \quad \dot{x}(0) = x_1. \end{cases}$$

A function $x \in C(0,T;V)$ is called a solution to the problem (E) if and only if $\dot{x} \in \mathcal{W}$ and (E) is satisfied.

We will need the following hypotheses.

$\underline{H(A)}$: $A\colon (0,T) \times V \to V^*$ is an operator such that

(1) $t \mapsto A(t,v)$ is measurable, for every $v \in V$,

(2) $v \mapsto A(t,v)$ is monotone and hemicontinuous, a.e. $t \in (0,T)$,

(3) $\langle A(t,v), v\rangle \geq c\,\|v\|^p - d\,|v|^2$ a.e. for all $v \in V$ with $c > 0$ and $d \geq 0$,

(4) $\|A(t,v)\|_{V^*} \leq a(t) + b\,\|v\|^{p-1}$ for all $v \in V$, a.e. $t \in (0,T)$ with $a \in L^q_+(0,T)$ and $b > 0$.

$\underline{H(B)}$: $B \in \mathcal{L}(V, V^*)$ is symmetric and coercive (i.e., $\langle Bv, v\rangle \geq m\,\|v\|^2$ for all $v \in V$ with $m > 0$).

$\underline{(H_0)}$: $x_0 \in V,\ x_1 \in H$.

The proof of the following result follows from the standard application of the Galerkin method and can be found in [1,5,6].

Proposition 1. *Under hypotheses $H(A)$, $H(B)$, (H_0) and $f \in \mathcal{H}^*$, the problem (E) admits a unique solution which satisfies $x \in C(0,T;V)$, $\dot{x} \in \mathcal{W}$, and the following estimate*

$$\|x(t)\|^2 + |\dot{x}(t)|^2 + \|\dot{x}\|^2_{\mathcal{W}} \leq C\left(1 + \|x_0\|^2 + |x_1|^2 + \|B\|^2_{\mathcal{L}(V,V^*)} + \|f\|^q_{\mathcal{H}^*}\right)$$

for all $t \in [0,T]$ with $C > 0$.

We present now a result on the continuous dependence of solutions to the problem

$$\text{(E)}_n \qquad \begin{cases} \ddot{x}(t) + A_n(t, \dot{x}(t)) + B_n x(t) = f_n(t) \ \text{ a.e. } t \in (0,T), \\ x(0) = x_0^n, \quad \dot{x}(0) = x_1^n. \end{cases}$$

on the data. We will need the following assumptions.

$\underline{H(A)_1}$: $A\colon (0,T) \times V \to V^*$ is such that $H(A)$ holds, $A_n\colon (0,T) \times V \to V^*$ satisfy $H(A)(1)(2)(3)$ uniformly with respect to $n \in \mathbb{N}$ and the condition

$$\|A_n(t,v)\|_{V^*} \leq a_n(t) + b\,\|v\|^{p-1} \quad \text{for all } v \in V, \text{ a.e. } t \in (0,T)$$
with $a_n \in L^q_+(0,T)$, $\sup_{n\in\mathbb{N}} ||a_n||_{L^q} < +\infty$, $b > 0$ and
$$A_n(\cdot, w(\cdot)) \to A(\cdot, w(\cdot)) \text{ in } s\text{–}\mathcal{V}^* \text{ for all } w \in \mathcal{V} \cap L^\infty(0,T;H).$$

$\underline{H(B)_1}$: $B_n \in \mathcal{L}(V,V^*)$ satisfy $H(B)$ uniformly with respect to $n \in \mathbb{N}$ and $B_n \to B$ in $\mathcal{L}(V,V^*)$.

$\underline{(H_0)_1}$: $x_0^n, x_0 \in V$, $x_1^n, x_1 \in H$, $x_0^n \to x_0$ in s–V and $x_1^n \to x_1$ in s–H.

For every $n \in \mathbb{N}$, let x_n be a solution of the problem $(\mathrm{E})_n$ and let x be a solution of the problem (E). We have

Proposition 2. *If hypotheses $H(A)_1$, $H(B)_1$, $(H_0)_1$ hold, $f_n \in \mathcal{H}^*$, $f_n \to f$ weakly in $\mathcal{H}^*$, then the sequence $\{(x_n, \dot{x}_n)\}$ converges to $(x, \dot{x})$ in $C(0,T;V\times H)$, as $n \to +\infty$.*

Proof. By Proposition 1 we know that, for every $n \in \mathbb{N}$, the problem $(\mathrm{E})_n$ has the unique solution $x_n \in C(0,T;V)$ such that $\dot{x}_n \in \mathcal{W}$. From $(\mathrm{E})_n$ and (E), we have

$$\langle \ddot{x}_n(s) - \ddot{x}(s), \dot{x}_n(s) - \dot{x}(s)\rangle + \langle A_n(s,\dot{x}_n(s)) - A(s,\dot{x}(s)), \dot{x}_n(s) - \dot{x}(s)\rangle +$$
$$+\langle B_n x_n(s) - Bx(s), \dot{x}_n(s) - \dot{x}(s)\rangle = \langle f_n(s) - f(s), \dot{x}_n(s) - \dot{x}(s)\rangle \quad \text{a.e.}$$

for every $n \in \mathbb{N}$. Integrating this equality and using the monotonicity of $A_n(s,\cdot)$, we get

$$|\dot{x}_n(t) - \dot{x}(t)|^2 - |\dot{x}_n(0) - \dot{x}(0)|^2 + 2\int_0^t \langle A_n(s,\dot{x}(s)) - A(s,\dot{x}(s)), \dot{x}_n(s) - \dot{x}(s)\rangle\, ds +$$
$$+2\int_0^t \langle B_n x_n(s) - Bx_n(s), \dot{x}_n(s) - \dot{x}(s)\rangle\, ds + \langle Bx_n(t) - Bx(t), x_n(t) - x(t)\rangle -$$
$$-\langle Bx_n(0) - Bx(0), x_n(0) - x(0)\rangle \leq 2\int_0^t \langle f_n(s) - f(s), \dot{x}_n(s) - \dot{x}(s)\rangle\, ds$$

for all $t \in [0,T]$. Hence using $H(B)_1$ and applying the Hölder inequality, we obtain

$$|\dot{x}_n(t) - \dot{x}(t)|^2 + m\,\|x_n(t) - x(t)\|^2 \leq \|B\|\,\|x_0^n - x_0\| + |x_1^n - x_1|^2 + \tag{1}$$
$$+2\,\|\widehat{A}_n(\dot{x}) - \widehat{A}(\dot{x})\|_{\mathcal{V}^*}\,\|\dot{x}_n - \dot{x}\|_{\mathcal{V}} + \widetilde{C}\,\|B_n - B\|\,\|x_n\|_{\mathcal{V}}\,\|\dot{x}_n - \dot{x}\|_{\mathcal{V}} +$$
$$+2\,\langle\langle f_n - f, \dot{x}_n - \dot{x}\rangle\rangle$$

for all $t \in [0,T]$, where $\widehat{A}_n$ and $\widehat{A}$ are the Nemitsky operators corresponding to A_n and A, respectively, and $\widetilde{C}$ is a positive constant independent of n. On the other hand, due to $H(A)_1$, $H(B)_1$ and $(H_0)_1$, from Proposition 1, we have

$$\|x_n(t)\|^2 + |\dot{x}_n(t)|^2 + \|\dot{x}_n\|^2_{\mathcal{W}} \leq C\left(1 + \|x_0^n\|^2 + |x_1^n|^2 + \|B_n\|^2 + \|f_n\|^q_{\mathcal{H}^*}\right). \tag{2}$$

Hence, it follows that $\{\dot{x}_n\}$ lies in a bounded subset of $\mathcal{W}$. Thus, up to a subsequence, $\dot{x}_n$ converges weakly in $\mathcal{W}$ and (since $\mathcal{W} \subset \mathcal{H}$ compactly) strongly in $\mathcal{H}$. So we have

$$\lim_{n \to +\infty} \langle\langle f_n - f, \dot{x}_n - \dot{x} \rangle\rangle = 0. \tag{3}$$

Using the assumptions, (2) and (3), from (1), we get $(x_n(t), \dot{x}_n(t)) \to (x(t), \dot{x}(t))$ in $s{-}(V \times H)$ for all $t \in [0,T]$, as $n \to +\infty$. Since the solution to (E) is unique, we deduce that the whole sequence $\{(x_n, \dot{x}_n)\}$ converges to $(x, \dot{x})$ in $C(0,T;V \times H)$. The proof is completed. □

In the sequel, we make use of the solution map $r\colon \mathcal{H}^* \to C(0,T;V) \times \mathcal{W}$ for (E) defined by $r(f) = (x, \dot{x})$, where x (respectively $\dot{x}$) is the solution (and its derivative, respectively) to (E). By Proposition 1 this map is well defined and Proposition 2 implies the following result.

Corollary 1. *Under hypotheses $H(A)$, $H(B)$ and (H_0), the solution map r for the problem (E) is continuous from $w{-}\mathcal{H}^*$ into $C(0,T;V \times H)$.*

3 Existence Result for Inclusions

In this section we study the existence of solutions to Problem (P). We start with the following

Definition 1. *A couple $(x,f) \in C(0,T;V) \times \mathcal{H}^*$ is called a mild solution to Problem* (P) *if and only if x is a solution to the evolution equation (E) and $f(\cdot) \in S^q_{F(\cdot,x(\cdot),\dot{x}(\cdot))}$.*

Prior to the existence theorem, we state the a priori bound on the solution to the evolution inclusion. We need the following hypotheses.

$H(F)$: $F\colon (0,T) \times H \times H \to \mathcal{P}_{fc}(H)$ is a multifunction such that

(1) F is graph measurable,
(2) $GrF(t,\cdot,\cdot)$ is sequentially closed in $H \times H \times (w{-}H)$, a.e. $t \in (0,T)$,
(3) $|F(t,x,y)| \le a_1(t) + b_1|x|^{2/q} + c_1|y|^{2/q}$, a.e. $t \in (0,T)$, where $a_1 \in L^q_+(0,T)$ and $b_1, c_1 > 0$.

Lemma 1. *Assume $H(A)$, $H(B)$, $H(F)$ and (H_0). If (x,f) is a mild solution to Problem* (P)*, then $(x,\dot{x},f)$ lies in a bounded set of $(L^\infty(0,T;V) \cap W^{1,\infty}(0,T;H)) \times \mathcal{W} \times \mathcal{H}^*$.*

In the proof of the next result, we follow methods used in [4,10].

Theorem 1. *If hypotheses $H(A)$, $H(B)$, $H(F)$ and (H_0) hold, then Problem* (P) *admits a mild solution.*

Proof. From Lemma 1, it is clear that every solution to Problem (P) satisfies

$$|x(t)| \le M_1, \qquad |\dot{x}(t)| \le M_2, \tag{4}$$

for all $t \in (0,T)$ with positive constants M_1, M_2. We define multifunction $\widehat{F}\colon (0,T) \times H \times H \to \mathcal{P}_{fc}(H)$ by $\widehat{F}(t,x,y) = F(t,p(x,y))$, where the map $p\colon H \times H \to B(0,M_1) \times B(0,M_2)$ is as follows

$$p(x,y) = \begin{cases} (x,y) & \text{if } |x| \le M_1 \text{ and } |y| \le M_2, \\ ((M_1 x/|x|),(M_2 y/|y|)) & \text{if } |x| > M_1 \text{ and } |y| > M_2, \\ ((M_1 x/|x|),y) & \text{if } |x| > M_1 \text{ and } |y| \le M_2, \\ (x,(M_2 y/|y|)) & \text{if } |x| \le M_1 \text{ and } |y| > M_2. \end{cases}$$

Since the map p is Lipschitz continuous, from the properties of F, we deduce that $\widehat{F}$ satisfies $H(F)(1)(2)$. Furthermore, we note that $|\widehat{F}(t,x,y)| \le \widetilde{a}_1(t)$ a.e. $t \in (0,T)$, where $\widetilde{a}_1 \in L^q_+(0,T)$ is given by $\widetilde{a}_1(t) = a_1(t) + b_1 M_1^{2/q} + c_1 M_2^{2/q}$.

We define $\mathcal{Z} = \{f \in \mathcal{H}^* \mid |f(t)| \le \widetilde{a}_1(t) \text{ a.e. } t \in (0,T)\}$ and a multifunction $\mathcal{R}$ on $\mathcal{Z}$ by

$$\mathcal{R}(f) = S^1_{\widehat{F}(\cdot,r(f)(\cdot))} = \left\{ f \in L^1(0,T;H) \mid f(t) \in \widehat{F}(t,r(f)(t)) \ \text{ a.e. } t \in (0,T) \right\}$$

(recall that $r(\cdot)$ is the solution map for the equation (E)). Since $\widehat{F}$ is graph measurable and L^1 integrably bounded, using the Aumann selection theorem (see Theorem 4.3.7 of [3]), we have $\mathcal{R}(f) \ne \emptyset$ for $f \in \mathcal{Z}$. Moreover, because $\widehat{F}$ is $\mathcal{P}_{fc}(H)$-valued and $|\widehat{F}(t,r(f)(t))| \le \widetilde{a}_1(t)$ a.e. $t \in (0,T)$, we obtain that $\mathcal{R}\colon \mathcal{Z} \to \mathcal{P}_{fc}(\mathcal{Z})$.

We will show that $\mathcal{R}$ is $(\mathrm{w}{-}\mathcal{H}^*){\times}(\mathrm{w}{-}\mathcal{H}^*)$ usc on $\mathcal{Z}$. Since $\mathcal{Z}$ is compact in $w{-}\mathcal{H}^*$, it suffices to prove (see Chapter I of [2], Sect. 4.1 of [3]) that $Gr\mathcal{R}$ is weakly$-$weakly closed in $\mathcal{Z} \times \mathcal{Z}$. Let $(f_n, z_n) \in Gr\mathcal{R}$, $f_n \to f$ and $z_n \to z$ both in $w{-}\mathcal{H}^*$. By Corollary 1, we know that $r(f_n)(t) \to r(f)(t)$ in $(s{-}H) \times (s{-}H)$ for all $t \in [0,T]$. Since $\widehat{F}$ satisfies $H(F)(1)(2)$, we deduce that $w{-}\limsup \widehat{F}(t,r(f_n)(t)) \subset \widehat{F}(t,r(f)(t))$ a.e. $t \in (0,T)$. Using Theorem 4.7.51 of [3], we obtain

$$w{-}\limsup \mathcal{R}(f_n) = w{-}\limsup S^1_{\widehat{F}(\cdot,r(f_n)(\cdot))} \subset$$

$$\subset S^1_{w-\limsup \widehat{F}(\cdot,r(f_n)(\cdot))} \subset S^1_{\widehat{F}(\cdot,r(f)(\cdot))} = \mathcal{R}(f).$$

From these inclusions we have $(f,z) \in Gr\mathcal{R}$. This means that $Gr\mathcal{R}$ is closed in $(w{-}\mathcal{Z}) \times (w{-}\mathcal{Z})$ and proves that $\mathcal{R}$ is weakly$-$weakly usc on $\mathcal{Z}$.

We apply the well known Kakutani-KyFan fixed point theorem for set-valued mappings (see Chapter I.12 of [2]) to the multifunction $\mathcal{R}$. We deduce that there exists $f^* \in \mathcal{Z}$ such that $f^* \in \mathcal{R}(f^*)$. The corresponding pair $(x^*,\dot{x}^*) = r(f^*)$ is a solution to Problem (P) with F replaced by $\widehat{F}$. However, the same estimates as in Lemma 1 (cf. also (4)), imply that $|x^*(t)| \le M_1$, $|\dot{x}^*(t)| \le M_2$ for every $t \in (0,T)$. Thus $\widehat{F}(t,x^*(t),\dot{x}^*(t)) = F(t,x^*(t),\dot{x}^*(t))$ for a.e. $t \in (0,T)$, which means that (x^*,f^*) is a mild solution to Problem (P). This completes the proof of the theorem. □

Corollary 2. *If $F(t,u,v) = \{f(t,u,v)\}$, where $f\colon (0,T)\times H\times H \to H$ is a function measurable in t, continuous in (u,v) and*

$$|f(t,u,v)| \le a_1(t) + b_1|u|^{2/q} + c_1|v|^{2/q} \quad \text{a.e. } t\in(0,T) \tag{5}$$

for all $u,v\in H$, then Theorem 1 ensures that the Cauchy problem for the nonlinear equation $\ddot{x}(t) + A(t,\dot{x}(t)) + Bx(t) = f(t,x(t),\dot{x}(t))$ has at least one solution.

Let S be the set of mild solutions to Problem (P) and let

$$\mathcal{M} = \{(x,\dot{x},f)\in C(0,T;V\times H)\times\mathcal{H}^* \mid (x,f)\in S\}.$$

Corollary 3. *Under the hypotheses of Theorem 1, the set $\mathcal{M}$ is nonempty, compact subset of $C(0,T;V\times H)\times(w{-}\mathcal{H}^*)$.*

Proof. The nonemptiness of $\mathcal{M}$ follows from Theorem 1. Let $\{(x_k,\dot{x}_k,f_k)\}_{k\in\mathbb{N}} \subset \mathcal{M}$. We will show that this sequence has a subsequence which converges in an appropriate topology to an element of $\mathcal{M}$. By the definition, x_k satisfies the evolution equation (E) with the right-hand side f_k and $f_k(\cdot)\in S^q_{F(\cdot,x_k(\cdot),\dot{x}_k(\cdot))}$. From Lemma 1, we obtain in particular that f_k remains in a bounded subset of $\mathcal{H}^*$. So after a possible passing to subsequence, we have $f_k \to f$ weakly in $\mathcal{H}^*$, as $k\to+\infty$, with $f\in\mathcal{H}^*$. Corollary 1 says that $r(f_k)\to r(f)$ in $C(0,T;V\times H)$, where $r(f)=(x,\dot{x})$ is a solution to (E). In order to conclude the proof, it suffices to show that f is a selection for $F(\cdot,x(\cdot),\dot{x}(\cdot))$. From Theorem 4.7.44 of [3], we have

$$f(t)\in \bar{co}\ w{-}\limsup\{f_k(t)\}_{k\ge1} \subset \bar{co}\ w{-}\limsup F(t,x_k(t),\dot{x}_k(t))$$

a.e. $t\in(0,T)$. Since $(x_k(t),\dot{x}_k(t)) \to (x(t),\dot{x}(t))$ in $s{-}(H\times H)$ for all $t\in[0,T]$, from $H(F)(1)$ (2), we easily deduce that $w{-}\limsup F(t,x_k(t),\dot{x}_k(t)) \subset F(t,x(t),\dot{x}(t))$ a.e. $t\in(0,T)$. Hence, we get $f(t)\in F(t,x(t),\dot{x}(t))$ a.e. $t\in(0,T)$. So we have obtained $(x,\dot{x},f)\in\mathcal{M}$ which completes the proof. □

4 Upper Semicontinuity Property of the Solution Set

Consider now a sequence of evolution inclusions Problem $(\mathrm{P})_\mathrm{n}$. Let us denote by S_n the set of mild solutions to Problem $(\mathrm{P})_\mathrm{n}$, i.e., $S_n = \{(x,f)\in C(0,T;V)\times\mathcal{H}^* \mid (x,f)$ is a mild solution to Problem $(\mathrm{P})_\mathrm{n}\}$.

Theorem 2. *Suppose that hypotheses $H(A)_1$, $H(B)_1$, $(H_0)_1$ hold, F, $F_n\colon (0,T)\times H\times H \to \mathcal{P}_{fc}(H)$ are multifunctions satisfying $H(F)$ uniformly with respect to $n\in\mathbb{N}$ and*

$$K\,(s{-}(H\times H)\times(w{-}H)) \limsup_{n\to+\infty,(u,v)\to(\tilde{u},\tilde{v})} F_n(t,u,v) \subset F(t,\tilde{u},\tilde{v}) \quad \text{a.e.} \tag{6}$$

If $(x_n,f_n)\in S_n$, $n\in\mathbb{N}$ and $f_n\to f$ in $w{-}\mathcal{H}^*$, then $(x,f)\in S$.

Proof. From Theorem 1 we know that S_n, $S \neq \emptyset$. Let $(x_n, f_n) \in S_n$ for $n \in \mathbb{N}$ and $f_n \to f$ weakly in $\mathcal{H}^*$. By Proposition 2, we infer that $(x_n, \dot{x}_n)$ converges in $C(0,T;V \times H)$ to $(x, \dot{x})$, as $n \to +\infty$, where x is a solution to the equation (E) (corresponding to the right hand side f). It remains to prove that $f(\cdot) \in S^q_{F(\cdot,x(\cdot),\dot{x}(\cdot))}$. From Theorem 4.7.44 of [3], we have

$$f(t) \in \bar{co}\ w\text{-}\limsup\{f_n(t)\}_{n\in\mathbb{N}} \subset \bar{co}\ w\text{-}\limsup F_n(t, x_n(t), \dot{x}_n(t))$$

a.e. $t \in (0,T)$ and by (6) we obtain

$$w\text{-}\limsup F_n(t, x_n(t), \dot{x}_n(t)) \subset F(t, x(t), \dot{x}(t)) \quad \text{a.e. } t \in (0,T).$$

This facts imply $f(t) \in F(t, x(t), \dot{x}(t))$ a.e. $t \in (0,T)$. Hence (x, f) is a mild solution to Problem (P) which concludes the proof. □

We introduce the sets $\mathcal{M}_n = \{(x, \dot{x}, f) \in C(0,T;V \times H) \times \mathcal{H}^* \mid (x,f) \in S_n\}$ for every $n \in \mathbb{N}$. We have the following upper semicontinuity property.

Corollary 4. *If hypotheses of Theorem 2 hold, then* $\limsup \mathcal{M}_n \subset \mathcal{M}$, *where the upper limit is taken in* $C(0,T;V \times H) \times (w\text{-}\mathcal{H}^*)$ *topology.*

5 Lower Semicontinuity Property of the Solution Set

In order to state a result on lower semicontinuity of the set of mild solutions, we admit the following stronger assumption on the multivalued term.

$\underline{H(F)_1}$: $F, F_n : (0,T) \times H \times H \to \mathcal{P}_{fc}(H)$ are multifunctions satisfying uniformly with respect to $n \in \mathbb{N}$ the conditions

(1) $F(\cdot, u, v)$ is measurable, for all $u, v \in H$,
(2) $F(t, \cdot, \cdot)$ is h-continuous, a.e. $t \in (0,T)$,
(3) $H(F)(3)$ holds

and

$$h(F_n(t,u_1,v_1), F(t,u_2,v_2)) \leq \alpha_n(t)\,(|u_1 - u_2| + |v_1 - v_2|) + \beta_n(t) \tag{7}$$

a.e. $t \in (0,T)$, with $\alpha_n \in L^1_+(0,T)$, $\alpha(t) = \sup_{n\in\mathbb{N}} \alpha_n(t) \in L^1_+(0,T)$ and $\beta_n \to 0$ in $L^2(0,T)$, as $n \to +\infty$.

Remark 1. The estimate (7) holds, for instance, if we suppose that

(a) $h(F_n(t,u_1,v_1), F_n(t,u_2,v_2)) \leq \alpha_n(t)\,(|u_1 - u_2| + |v_1 - v_2|)$ a.e., for every $n \in \mathbb{N}$, $u_1, u_2, v_1, v_2 \in H$,
(b) $F_n(t,u,v) \to F(t,u,v)$ in the Hausdorff metric, for all $u, v \in H$, a.e. t.

Theorem 3. *If hypotheses* $H(A)_1$, $H(B)_1$, $H(F)_1$ *and* $(H_0)_1$ *hold, then* $\mathcal{M} \subset \liminf \mathcal{M}_n$, *where the lower limit is taken in* $C(0,T;V \times H) \times (s\text{-}\mathcal{H}^*)$ *topology.*

Proof. Let $(x, \dot{x}, f) \in \mathcal{M}$. We have to find $(x_n, \dot{x}_n, f_n) \in \mathcal{M}_n$ such that

$$(x_n, \dot{x}_n) \to (x, \dot{x}) \ \text{ in } \ C(0,T; V \times H), \tag{8}$$

$$f_n \to f \ \text{ in } \ s\text{–}\mathcal{H}^*. \tag{9}$$

Define $f_n(t,u,v) = \text{proj}(f(t), F_n(t,u,v))$ for $n \in \mathbb{N}$, where $\text{proj}(a, \mathcal{A})$ denotes the projection of point a onto the set $\mathcal{A}$. Due to Lemma α of [10], we have that f_n is measurable in t and continuous in (u,v). Moreover, $f_n(t,u,v) \in F_n(t,u,v)$ and $H(F)(3)$ implies that f_n satisfies the growth condition (5). Therefore applying Corollary 2, we obtain that for every $n \in \mathbb{N}$, the problem

$$\begin{cases} \ddot{x}(t) + A_n(t, \dot{x}(t)) + B_n x(t) = \{f_n(t, x(t), \dot{x}(t))\} \ \text{ a.e. } t \in (0,T), \\ x(0) = x_0, \ \ \dot{x}(0) = x_1 \end{cases}$$

possesses a solution $x_n \in C(0,T;V)$ with $\dot{x}_n \in \mathcal{W}$. From the equality

$$\langle \ddot{x}_n(s) - \ddot{x}(s), \dot{x}_n(s) - \dot{x}(s) \rangle + \langle A_n(s, \dot{x}_n(s)) - A(s, \dot{x}(s)), \dot{x}_n(s) - \dot{x}(s) \rangle +$$

$$+ \langle B_n x_n(s) - Bx(s), \dot{x}_n(s) - \dot{x}(s) \rangle = \langle f_n(s, x_n(s), \dot{x}_n(s)) - f(s), \dot{x}_n(s) - \dot{x}(s) \rangle$$

a.e. $s \in (0,T)$, by integrating by parts, using $H(A)_1$, $H(B)_1$, similarly as in the proof of Proposition 2, we obtain

$$|\dot{x}_n(t) - \dot{x}(t)|^2 + m \, \|x_n(t) - x(t)\|^2 \le \sigma_n +$$

$$+ 2 \int_0^t |f_n(s, x_n(s), \dot{x}_n(s)) - f(s)| |\dot{x}_n(s) - \dot{x}(s)| \, ds$$

for all $t \in [0,T]$, where $\sigma_n = 2 \, \|\widehat{A}_n(\dot{x}) - \widehat{A}(\dot{x})\|_{\mathcal{V}^*} \|\dot{x}_n - \dot{x}\|_{\mathcal{V}} + C \, \|B_n - B\| \, \|x_n\|_{\mathcal{V}}$ $\|\dot{x}_n - \dot{x}\|_{\mathcal{V}}$ and $C > 0$. Taking into account that

$$|f_n(s, x_n(s), \dot{x}_n(s)) - f(s)| = d\,(f(s), F_n(s, x_n(s), \dot{x}_n(s))) \le \tag{10}$$

$$\le h\,(F(s, x(s), \dot{x}(s)), F_n(s, x_n(s), \dot{x}_n(s))) \le$$

$$\le \alpha_n(s)\,(|x_n(s) - x(s)| + |\dot{x}_n(s) - \dot{x}(s)|) + \beta_n(s) \ \text{ a.e. } s \in (0,T),$$

we have

$$|\dot{x}_n(t) - \dot{x}(t)|^2 + m \, \|x_n(t) - x(t)\|^2 \le \sigma_n + 2 \int_0^t \alpha(s) |\dot{x}_n(s) - \dot{x}(s)|^2 \, ds +$$

$$+ 2 \int_0^t \alpha(s) |x_n(s) - x(s)| \, |\dot{x}_n(s) - \dot{x}(s)|^2 \, ds + 2 \int_0^t \beta_n(s) |\dot{x}_n(s) - \dot{x}(s)| \, ds$$

for all $t \in [0,T]$. Applying the inequality $2ab \le a^2 + b^2$, $a, b > 0$ to the last two integrals on the right hand side and using the fact that $|\cdot| \le \gamma \|\cdot\|$ with $\gamma > 0$, we have

$$|\dot{x}_n(t) - \dot{x}(t)|^2 + m \, \|x_n(t) - x(t)\|^2 \le \sigma_n + \|\beta_n\|^2_{L^2(0,T)} +$$

$$+\int_0^t \left[(3\alpha(s)+1)\,|\dot{x}_n(s)-\dot{x}(s)|^2+\alpha(s)\gamma^2\|x_n(s)-x(s)\|^2\right]\,ds$$

for all $t\in[0,T]$. Invoking the Gronwall inequality, we get

$$|\dot{x}_n(t)-\dot{x}(t)|^2+\|x_n(t)-x(t)\|^2\le C\left(\sigma_n+\|\beta_n\|^2_{L^2}\right)\quad\text{for all } t\in(0,T),$$

where C is a positive constant independent of n. From Lemma 1, $H(A)_1$ and $H(B)_1$, we infer that $\lim\sigma_n=0$. Hence, we have shown (8).

In order to prove (9), by using (10), we write

$$\int_0^T |f_n(s,x_n(s),\dot{x}_n(s))-f(s)|^q\,ds\le$$

$$\le 2^{q-1}\int_0^T(\alpha(s))^q(|x_n(s)-x(s)|+|\dot{x}_n(s)-\dot{x}(s)|)^q\,ds+2^{q-1}\|\beta_n\|^2_{L^2}.$$

In view of (8) and the convergence $\beta_n\to 0$ in $L^2(0,T)$, we easily get (9). This completes the proof. □

Corollary 5. *If hypotheses of Theorem 3 hold, then* $\mathcal{M}_n\xrightarrow{K}\mathcal{M}$ *in* $C(0,T;V\times H)\times(s\text{–}\mathcal{H}^*)$ *topology. This follows from Theorem 3 and the fact that Corollary 4 implies* $\limsup\mathcal{M}_n\subset\mathcal{M}$ *in this topology.*

References

1. Ahmed, N.U., Kerbal, S.: Optimal control of nonlinear second order evolution equations. J. Appl. Math. Stochastic Anal. **6**, 123–136 (1993)
2. Aubin, J.-P., Cellina, A.: Differential Inclusions. Springer, New York (1984)
3. Denkowski, Z., Migórski, S., Papageorgiou, N.S.: An Introduction to Nonlinear Analysis: Theory. Kluwer Academic/Plenum Publishers, New York (2003)
4. Denkowski, Z., Mortola, S.: Asymptotic behavior of optimal solutions to control problems for systems described by differential inclusions corresponding to partial differential equations. J. Optimiz. Theory Appl. **78**, 365–391 (1993)
5. Lions, J.L.: Optimal Control of Systems Governed by Partial Differential Equations. Springer, New York (1971)
6. Migórski, S.: Variational stability analysis of optimal control problems for systems governed by nonlinear second order evolution equations. J. Math. Syst. Estim. Control **6**, 1–24 (1996)
7. Migórski, S.: Control problems for systems described by nonlinear second order evolution inclusions. Nonlinear Anal. Theory Meth. Appl. **30**, 419–428 (1997)
8. Migórski, S.: Existence of solutions to nonlinear second order evolution inclusions without and with impulses. Dyn. Cont. Discr. Impul. Syst. Ser. B **18**, 493–520 (2011)
9. Migórski, S., Ochal, A., Sofonea, M.: Nonlinear Inclusions and Hemivariational Inequalities. Models and Analysis of Contact Problems, Advances in Mechanics and Mathematics, vol. 26. Springer, New York (2013)
10. Papageorgiou, N.S.: Continuous dependence results for a class of evolution inclusions. Proc. Edinburgh Math. Soc. **35**, 139–158 (1992)
11. Zeidler, E.: Nonlinear Functional Analysis and its Applications II. Springer, New York (1990)

Impulse Control of Standard Brownian Motion: Long-Term Average Criterion

Kurt Helmes[1], Richard H. Stockbridge[2(✉)], and Chao Zhu[2]

[1] Institut für Operations Research, Humboldt-Universität zu Berlin, Berlin, Germany
helmes@wiwi.hu-berlin.de
[2] Department of Mathematical Sciences, University of Wisconsin – Milwaukee, Milwaukee, WI 53201, USA
{stockbri,zhu}@uwm.edu

Abstract. This paper examines the impulse control of a standard Brownian motion under a long-term average criterion. In contrast with the dynamic programming approach, this paper first imbeds the stochastic control problem into an infinite-dimensional linear program over a space of measures and then reduces the problem to a simpler nonlinear optimization that has a familiar interpretation. One is able to easily identify the optimal cost and a family of optimal impulse control policies.

Keywords: Impulse control · Long-term average criterion · Infinite dimensional linear programming · Expected occupation and impulse measures

1 Introduction

When one seeks to control a stochastic process and every intervention incurs a strictly positive cost, one must select a sequence of separate intervention times and amounts. The resulting stochastic problem is therefore an impulse control problem in which the decision maker seeks to either maximize a reward or minimize a cost. This paper examines the impulse control of the prototypical process Brownian motion under a long-term average cost criterion; a companion paper studies the impulse control of Brownian motion under a discounted criterion. The aim of the paper is to illustrate a solution approach which first imbeds the stochastic control problem into an infinite-dimensional linear program over a space of measures and then reduces the linear program to a simpler nonlinear optimization. This approach provides a new method for determining an optimal impulse control policy.

Let W be a standard Brownian motion process with natural filtration $\{\mathcal{F}_t\}$. An impulse control policy consists of a pair of sequences $(\tau, Y) := \{(\tau_k, Y_k) : k \in \mathbb{N}\}$ in which τ_k is the $\{\mathcal{F}_t\}$-stopping time of the kth impulseand the

This research was supported in part by National Science Foundation under grant DMS-1108782 and by grant award 246271 from the Simons Foundation.

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 148–157, 2014.
DOI: 10.1007/978-3-662-45504-3_14

$\mathcal{F}_{\tau_k}$-measurable variable Y_k gives the kth impulse size. The sequence $\{\tau_k : k \in \mathbb{N}\}$ is required to be non-decreasing, a natural assumption in that intervention $k+1$ must occur no earlier than intervention k. For a policy (τ, Y), the impulse-controlled Brownian motion process is given by

$$X(t) = x_0 + W(t) + \sum_{k=1}^{\infty} I_{\{\tau_k \le t\}} Y_k.$$

The goal of the decision maker is to keep the process close to zero and to minimize the long-term average cost incurred by the impulse policy. Define the running/deviation cost rate function by $c_0(x) = x^2$ and set the impulse costs to be $c_1(y, z) = k_1 + k_2|y - z|$ for $(y, z) \in \mathbb{R}^2$, in which $k_1 > 0$ and $k_2 \ge 0$ and y denotes the pre-intervention location which will typically be far away from zero, while z denotes the post-intervention location of the process X, which should be close to zero. Note, in particular, that there is a strictly positive cost for every impulse, even one in which $Y_k = 0$ which does not affect the value of X. Let (τ, Y) be an impulse control policy. The quantity to be minimized is

$$J(\tau, Y) := \limsup_{t\to\infty} t^{-1}\mathbb{E}\left[\int_0^t c_0(X(s))\,ds + \sum_{k=1}^{\infty} I_{\{\tau_k \le t\}} c_1(X(\tau_k-), X(\tau_k))\right]. \quad (1)$$

Clearly any policy (τ, Y) for which $J(\tau, Y) = \infty$ is undesirable so to be *admissible,* we require (τ, Y) to have a finite cost; the nonnegativity of c_0 and strict positivity of c_1 indicates that every policy will have nonnegative long-term average cost. The collection of all admissible impulse policies is denoted by $\mathcal{A}$.

Similar type of problems have been extensively investigated in the literature. An incomplete list includes the now classical works on general stochastic impulse problems [1,4,8,13] as well as their applications in various areas such as portfolio optimization, inventory control, risk management, control of a dam and exchange rate intervention [2,3,10–12]. In particular, [11] explains the adoption of a Brownian motion model.

Unlike the aforementioned references, in which the primary tool is the dynamic programming principle and its associated quasi-variational inequalities, this paper aims to illustrate the utility of a different methodology, namely, the linear programming approach. In such an approach, we embed the stochastic impulse control problem into an infinite-dimensional linear program over appropriate measures. Further, the linear program, with the aid of an auxiliary linear program, is transformed into a nonlinear optimization problem. Then both the value of the impulse control problem and an optimal impulse control policy are easily determined. The linear programming approach toward stochastic control problem can be dated back to [9] for discrete time and a finite state space and to [14,15] for regular stochastic control problems in continuous time with general state space. It has been further developed in [5–7] for optimal stopping and singular control problems. This paper aims to expand the utility of such a methodology to impulse control problems as well.

We make four important observations about impulse policies. Firstly, the "no-intervention policy" in which the process is $X(t) = x_0 + W(t)$ for all $t \geq 0$ incurs an infinite long-term average cost so is inadmissible. Secondly, let (τ, Y) be an impulse policy and define $\tau_\infty := \lim_{k\to\infty} \tau_k$. Should $\tau_\infty < \infty$ on a set of positive probability, then the fixed cost $k_1 > 0$ per intervention also results in an infinite long-term average cost. Thus for every admissible policy $\tau_k \to \infty$ a.s. as $k \to \infty$. Thirdly, let (τ, Y) be a policy for which there is some k such that $\tau_k = \tau_{k+1}$ on a set of positive probability. Again due to the presence of the fixed intervention cost k_1, the total cost up to time τ_{k+1} will be at least $k_1\mathbb{P}(\tau_k = \tau_{k+1})$ smaller by combining these interventions into a single intervention on this set. Hence we may restrict policies to those for which $\tau_k < \tau_{k+1}$ almost surely for each k.

The final observation is similar. Suppose (τ, Y) is a policy such that on a set G of positive probability $\tau_k < \infty$ and $|X(\tau_k)| > |X(\tau_k-)|$ for some k. Consider a modification of this impulse policy and resulting process $\tilde{X}$ which simply fails to implement this impulse on G. Define the stopping time $\sigma = \inf\{t > \tau_k : |X(t)| \leq |\tilde{X}(t)|\}$. Notice that the running costs accrued by $\tilde{X}$ over $[\tau_k, \sigma)$ are smaller than those accrued by X. Finally, at time σ, introduce an intervention on the set G which moves the $\tilde{X}$ process so that $\tilde{X}(\sigma) = X(\sigma)$. This intervention will incur a cost which is no greater than the cost for the process X at time τ_k. As a result, we may restrict the impulse control policies to those for which no impulse increases the distance of the process from the origin (an intuitively obvious observation).

2 Restricted Problem and Measure Formulation

The initial analysis considers a restricted collection of impulse policies.

Condition 1. *Let $\mathcal{A}_1$ denote the set of policies such that for $(\tau, Y) \in \mathcal{A}_1$ the resulting process X remains bounded; that is for some $M < \infty$, $|X(t)| \leq M$ for all $t \geq 0$.*

Intuitively, Condition 1 is not much of a restriction since unbounded processes occur by allowing the Brownian motion to diffuse which incurs an expensive running cost. This restriction, however, is needed so that a transversality condition is satisfied and a stochastic integral is a martingale. Following the initial solution, the general class of impulse policies will be analyzed.

We capture the expected behavior of the process and impulses with measures. Arbitrarily fix $(\tau, Y) \in \mathcal{A}_1$ and let M be as given in Condition 1. For each $t > 0$, define the average expected occupation and average expected impulse measures $\mu_0^{(t)}$ and $\mu_1^{(t)}$, respectively, such that for each $G, G_1, G_2 \in \mathcal{B}(\mathbb{R})$,

$$\mu_0^{(t)}(G) = t^{-1}\mathbb{E}\left[\int_0^t I_G(X(s))\,ds\right], \text{ and} \tag{2}$$

$$\mu_1^{(t)}(G_1 \times G_2) = t^{-1}\mathbb{E}\left[\sum_{k=1}^{\infty} I_{\{\tau_k \leq t\}} I_{G_1\times G_2}(X(\tau_k-), X(\tau_k))\right].$$

It is immediate that $\mu_0^{(t)}$ is a probability measure for each $t > 0$ and since $(\tau, Y) \in \mathcal{A}_1 \subset \mathcal{A}$, the finiteness of $J(\tau, Y)$ implies that the collection of measures $\{\mu_1^{(t)} : t > 0\}$ is uniformly bounded above. We also note that the second component being measured by $\mu_1^{(t)}$ is the post-jump location so $\mu_1^{(t)}$ is a measure on the product space of (pre-jump, post-jump) pairs. In light of Condition 1, the support of $\mu_1^{(t)}$ is contained in the compact set $\mathcal{R} := \{(y, z) : |z| \le |y| \le M\}$. Similarly, each $\mu_0^{(t)}$ has support in the compact interval $[-M, M]$. As a result, these collections are tight and hence relatively compact.

Now notice the objective function (1) can be expressed as

$$J(\tau, Y) = \limsup_{t\to\infty} t^{-1} \left[\int c_0(x)\, \mu_0^{(t)}(dx) + \int c_1(y, z)\, \mu_1^{(t)}(dy \times dz) \right].$$

Let $\{t_j : j \in \mathbb{N}\}$ be a sequence with $t_j \to \infty$ as $j \to \infty$ such that

$$t_j^{-1} \left[\int c_0(x)\, \mu_0^{(t_j)}(dx) + \int c_1(y, z)\, \mu_1^{(t_j)}(dy \times dz) \right] \to J(\tau, Y).$$

For $i = 0, 1$, the relative compactness of $\{\mu_i^{(t_j)} : j \in \mathbb{N}\}$ implies that there exist weak limits μ_0 and μ_1. Note μ_0 is a probability measure whereas μ_1 is a finite measure. Since c_0 and c_1 are bounded and continuous on $[-M, M]$ and $\mathcal{R}$, respectively,

$$J(\tau, Y) = \int c_0(x)\, \mu_0(dx) + \int c_1(y, z)\, \mu_1(dy \times dz). \tag{3}$$

It is now helpful to characterize the value of functions of the process. For $f \in \mathcal{D} = C^2(\mathbb{R})$,

$$\begin{aligned} f(X(t)) = f(x_0) &+ \int_0^t f'(X(s))\, dW(s) \\ &+ \int_0^t \frac{1}{2} f''(X(s))\, ds + \sum_{k=1}^{\infty} [f(X(\tau_k)) - f(X(\tau_k-))]\, I_{\{\tau_k \le t\}}. \end{aligned} \tag{4}$$

The generator A of the Brownian motion process is $Af(x) = \frac{1}{2} f''(x)$; define the jump operator B by $Bf(y, z) = f(z) - f(y)$. First taking expectations, then dividing by t and letting $t \to \infty$ in (4) results in

$$\limsup_{t\to\infty} t^{-1} \mathbb{E} \left[\int_0^t Af(X(s))\, ds + \sum_{k=1}^{\infty} I_{\{\tau_k \le t\}} Bf(X(\tau_k-), X(\tau_k)) \right] = 0; \tag{5}$$

note that the boundedness of $X(t)$ along with $f \in C^2(\mathbb{R})$ implies both the transversality condition $\lim_{t\to\infty} t^{-1}\mathbb{E}[f(X(t))] = 0$ holds and that the stochastic integral exists and has mean 0. (The same argument applies by taking the limit inferior in (4), so in fact, left-hand side of (5) is a limit.) The fact that f and f'' are continuous and bounded on $[-M, M]$ means that

$$\int Af(x)\, \mu_0^{(t_k)} \to \int Af(x)\, \mu_0(dx)$$

and

$$\int Bf(y,z)\,\mu_1^{(t_k)}(dy\times dz)\to\int B(y,z)\,\mu_1(dy\times dz)$$

and hence (5) can be written in terms of these measures as

$$\int Af(x)\,\mu_0(dx)+\int Bf(y,z)\,\mu_1(dy\times dz)=0. \tag{6}$$

The restricted impulse control problem is therefore imbedded in the linear program of minimizing (3) over pairs of measures (μ_0,μ_1) satisfying the constraints (6) for every $f\in\mathcal{D}$. Since there may be pairs (μ_0,μ_1) which do not correspond to any $(\tau,Y)\in\mathcal{A}_1$, it follows that the value of the linear program is a lower bound on the minimal long-run average cost of the impulse control problem. Observe also that by further restricting the collection of functions for which the constraint is required to be satisfied, the corresponding "auxiliary" linear program may have even more feasible pairs and hence will provide an even lower bound on the value of the impulse control problem. These observations are summarized in the following theorem.

Theorem 2. *Let V denote the optimal value of the long-term average impulse control problem, V_{lp} denote the optimal value of the linear program which seeks to minimize (3) over measures satisfying (6) and V_{aux} be the optimal value of an auxiliary linear program which limits (6) to a smaller collection of functions $f\in\mathcal{D}_1$ for some $\mathcal{D}_1\subset\mathcal{D}$. Then $V_{aux}\le V_{lp}\le V$.*

3 Partial Solution: First Auxiliary Linear Program

It would be helpful to reduce the complexity of the linear program by reducing the number of constraints. We first consider the constraints (6). The intuition is rather straightforward. Consider a function ϕ for which $A\phi\equiv-1$. Then since μ_0 is a probability measure for each feasible pair (μ_0,μ_1), the identity (6) becomes

$$\int_{\{|z|\le|y|\}}B\phi(y,z)\,\mu_1(dy\times dz)=1. \tag{7}$$

The general solution to the equation $A\phi\equiv-1$ is $\phi(x)=-x^2+ax+b$, in which a,b are constants. We shall select a solution ϕ so that $B\phi(x,y)=\mathbb{E}_x[\tau_y]$, where $|x|<|y|$ and $\tau_y:=\inf\{t\ge0:|X(t)|=|x+W(t)|=|y|\}$. To this end, we notice that the function $u(x):=-x^2+y^2$ solves the boundary value problem

$$\begin{cases}Au(x)=-1, & x\in(-|y|,|y|),\\ u(-|y|)=u(|y|)=0.\end{cases} \tag{8}$$

Therefore the optional sampling theorem implies that for any $x\in(-|y|,|y|)$ and $t\ge0$, we have

$$\mathbb{E}_x\left[u(X(t\wedge\tau_y))\right]=u(x)+\mathbb{E}_x\left[\int_0^{t\wedge\tau_y}Au(X(s))\,ds\right]=u(x)-\mathbb{E}_x[t\wedge\tau_y].$$

Since $X(t \wedge \tau_y)$ is bounded and $\tau_y < \infty$ a.s., utilizing the boundary conditions in (8), letting $t \to \infty$ in the above equation yields $\mathbb{E}_x[\tau_y] = u(x) = -x^2 + y^2$. Therefore by selecting $\phi(x) = -x^2, x \in \mathbb{R}$, the term $B\phi(x, y)$ in (7) gives the expected time it takes the Brownian motion process starting at x to hit the set $\{-|y|, |y|\}$.

In a similar manner, by considering the boundary value problem

$$\begin{cases} Au(x) = -c_0(x), \quad x \in (-|y|, |y|), \\ u(-|y|) = u(|y|) = 0, \end{cases} \tag{9}$$

it follows that the function $g_0(x) := -\frac{1}{6}\, x^4$ satisfies

$$Bg_0(x, y) = g_0(y) - g_0(x) = \mathbb{E}_x \left[\int_0^{\tau_y} c_0(X(s)) ds \right], \quad |x| < |y|. \tag{10}$$

The following proposition establishes the required identity.

Proposition 3. *Let $(\tau, Y) \in \mathcal{A}_1$ and let X denote the resulting impulse controlled process. Recall $\{t_j : j \in \mathbb{N}\}$ is a set of times such that*

$$t_j^{-1} \left[\int c_0 \, d\mu_0^{(t_j)} + \int c_1 \, d\mu_1^{(t_j)} \right] \to J(\tau, Y).$$

Let (μ_0, μ_1) be a weak limit of $(\mu_0^{(t_j)}, \mu_1^{(t_j)})$ as $j \to \infty$. Then

$$\int c_0(x)\, \mu_0(dx) = \int Bg_0(y, z)\, \mu_1(dy \times dz). \tag{11}$$

Proof. Without loss of generality, assume that $\mu_0^{(t_j)} \Rightarrow \mu_0$ and similarly $\mu_1^{(t_j)} \Rightarrow \mu_1$. Using g_0 in (4) and taking expectations yields for each t_j

$$\mathbb{E}_{x_0}[g_0(X(t_j))] = g_0(x_0) + \mathbb{E}_{x_0} \left[\int_0^{t_j} Ag_0(X(s))\, ds + \sum_{k-0}^{\infty} I_{\{\tau_k \le t_j\}} Bg_0(X(\tau_k-), X(\tau_k)) \right].$$

Since for $(\tau, Y) \in \mathcal{A}_1$, $X(t)$ remains bounded, dividing by t_j, using the definitions of $\mu_0^{(t_j)}$ and $\mu_1^{(t_j)}$ in (2) and letting $j \to \infty$ establishes the result. □

We are now ready to define the first auxiliary linear program. Restrict the constraint to the single function $\phi(x) = -x^2$ and use Proposition 3 to rewrite the objective function. The resulting linear program is

$$\begin{cases} \text{Min.} \displaystyle\int [c_1(y, z) + Bg_0(y, z)]\, \mu_1(dy \times dz) \\ \text{S.t.} \quad \displaystyle\int B\phi(y, z)\, \mu_1(dy \times dz) = 1. \end{cases} \tag{12}$$

Notice the support of each feasible μ_1 is in the set $\{(y,z) \in \mathbb{R}^2 : |z| \leq |y|\}$. Thus $B\phi(y,z) = y^2 - z^2 \geq 0$ and the constraint of (12) implies that $B\phi(y,z) = y^2 - z^2$ is a probability density for every feasible measure μ_1 of (12).

We emphasize that the auxiliary linear program includes the cost of any impulse policy $(\tau, Y) \in \mathcal{A}_1$. These policies are not required to be of feedback type or even stationary. It is only required that the resulting controlled process remain bounded.

Observe that the constraint does not impose any mass restrictions on the set $\{(y,z) : B\psi(y,z) = 0\}$. However, the goal is to minimize the objective function and since the impulse cost function $c_1 > k_1 > 0$ any mass placed on this set will only increase the cost. We may therefore restrict the optimization to those measures μ_1 having support in $\{(y,z) : B\phi(y,z) > 0\} = \{(y,z) : |z| < |y|\}$. As a result the objective function can be rewritten as

$$\begin{aligned} &\int [c_1(y,z) + Bg_0(y,z)]\, \mu_1(dy \times dz) \\ &\quad = \int \left(\frac{c_1(y,z) + Bg_0(y,z)}{B\phi(y,z)} \right) \cdot B\phi(y,z)\, \mu_1(dy \times dz) \end{aligned}$$

and so the problem reduces to the minimization of

$$F(y,z) = \frac{c_1(y,z) + Bg_0(y,z)}{B\phi(y,z)} \tag{13}$$

over $\{(y,z) : |z| \leq |y|\}$. Observe that $F(-y,-z) = F(y,z)$.

Remark 4. *The minimization of F in (13) has a very familiar interpretation. Let y and z be such that $0 \leq |z| < |y|$. Let $\tau_y = \inf\{t \geq 0 : |X(t)| = |y|\}$. Recall from (10) that $Bg_0(y,z)$ represents the expected running cost for the cycle $[0, \tau_y)$. Now consider the impulse policy in which impulses occur only when the process X hits either y or $-y$ and then jumps to z or $-z$, respectively. The symmetry of the fixed cost function c_1 means that $c_1(y,z) = c_1(-y,-z)$ and this cost is assessed at the end of the cycle. Note also $B\phi(y,z)$ gives the expected time it takes the Brownian motion starting from $-z$ or z to reach $-y$ or y. Hence the function F represents the ratio of the expected cost per cycle over the expected cycle length taken for such impulse control policies. The significance of this reformulation is that the linear programming imbedding allows arbitrary impulse policies in the class $\mathcal{A}_1$ yet the resulting nonlinear optimization corresponds to minimizing the cost over a subclass of these policies.*

We now determine an optimal impulse control policy.

Theorem 5. *There exist values $y_* > z_* > 0$ such that*

$$F(y_*, z_*) = F(-y_*, -z_*) = \inf_{(y,z) \in \mathcal{R}} F(y,z).$$

Proof. First since $B\phi(y,z) = y^2 - z^2$ and $Bg_0(y,z) = \frac{1}{6}(y^4 - z^4)$, the function $F(y,z) = \frac{k_1+k_2|y-z|}{y^2-z^2} + \frac{1}{6}(y^2+z^2)$. Observe that as $y \to \infty$ or $z \to \infty$ or $|y-z| \to 0$, $F(y,z) \to \infty$. Hence F achieves its minimal value at some point (y_*, z_*) which by optimality requires y_* and z_* to have the same sign. □

Theorem 6. *An optimizing pair (y_*, z_*) having positive components is the level set of the function $h(x) = \frac{k_2}{2x} + \frac{x^2}{3}$ at the level $F(y_*, z_*)$.*

Proof. Consider the function F on the set $\{(y,z) : 0 < z < y\}$ so $|y-z| = y - z$. The first-order optimality conditions are

$$\frac{[y^2 - z^2](k_2 + (2/3)y^3) - [k_1 + k_2(y-z) + (1/6)(y^4 - z^4)](2y)}{[y^2 - z^2]^2} = 0, \tag{14}$$

$$\frac{[y^2 - z^2](-k_2 - (2/3)z^3) + [k_1 + k_2(y-z) + (1/6)(y^4 - z^4)](2z)}{[y^2 - z^2]^2} = 0, \tag{15}$$

which are satisfied for pairs (y,z) such that

$$F(y,z) = h(y) = h(z), \tag{16}$$

where the function h is defined in the statement of the theorem. We have $h'(x) = \frac{4x^3 - 3k_2}{6x^2}$. When $x > 0$, one observes that h strictly decreases from $+\infty$ until it reaches a minimum at $\sqrt[3]{3k_2/4}$ after which it strictly increases to $+\infty$. For $x < 0$, h is strictly decreasing from $+\infty$ to $-\infty$. The level sets of h consist of either a single value with $x < 0$ when the level lies below the minimum of h over the positive reals or three points with $x < 0 < z < y$ when the level is above this minimum value. □

It is still necessary to connect the optimal value of the linear program (12) to the optimal value of the long-term average impulse control problem. Examine the relative magnitudes of the roots given in the proof of Theorem 6; by definition $0 < z < y$. Observe the root $x < 0$ is such that $|x| > y$ and hence $x < -y < -z$.

Theorem 7. *The optimal long-term average cost for the restricted impulse control problem is $F(y_*, z_*)$ and an optimal impulse control policy (τ^*, Y^*) is defined by:*

$$\begin{cases} \tau_1^* = 0, \\ Y_1^* = z_* - x_0, \end{cases} \qquad \begin{cases} \tau_k^* = \inf\{t > \tau_{k-1}^* : X(t-) = \pm y_*\}, \\ Y_k^* = \operatorname{sgn}(X(\tau_k^*-))z_* - X(\tau_k^*-), \end{cases} \quad k \geq 2. \tag{17}$$

Remark 8. *Due to the nature of the long-term average criterion, many other optimal policies exist. In fact, any impulse control policy $(\tau, Y) \in \mathcal{A}_1$ can be used for a finite length of time so long as after some point the process is in the interval $[-y_*, y_*]$ and the policy of jumping to z_* when the process hits y_* and jumping to $-z_*$ at the time of hitting $-y_*$ is adopted.*

Proof. Consider the impulse control policy (17) and observe that $(\tau^*, Y^*) \in \mathcal{A}_1$. Also for each $t > 0$ by (2), $\mu_1^{(t)}$ has its support on the set

$$\{(x_0, z_*), (-y_*, -z_*), (y_*, z_*)\}$$

but the limiting measure μ_1 only has mass on $\{(-y_*, -z_*), (y_*, z_*)\}$. As a result, $J(\tau^*, Y^*) = F(y_*, z_*)$ and hence by Theorem 2 is optimal. □

4 General Solution

The solution obtained in the previous section does not include impulse control policies which allow the process X to become unbounded in either direction. It is therefore necessary to show that such policies cannot provide a lower cost.

Theorem 9. *The optimal long-term average cost over the collection of all admissible impulse control policies is $F(y_*, z_*)$ and any policy which eventually only impulses when the process reaches $\pm y_*$, with impulses respectively to $\pm z_*$, is optimal.*

Proof. Let (τ, Y) be an arbitrary admissible impulse control policy in $\mathcal{A}$ and let X denote the resulting process. We shall prove that F_* is a lower bound on $J(\tau, Y)$ and since F_* is achieved by the impulse control policy (17), this will establish the result.

We use a localization argument. For each $n \in \mathbb{N}$, define the stopping time $\sigma_n = \inf\{t \geq 0 : |X(t)| \geq n\}$ and to simplify notation, let $F_* = F(y_*, z_*)$. We consider the function $f(x) = F_*\phi(x) - g_0(x) = \frac{1}{6}x^4 - F_*x^2$, $x \in \mathbb{R}$. Due to the choices of ϕ and g_0, we have $Af(x) = c_0(x) - F_*$, and for any $|z| < |y|$,

$$\begin{aligned} Bf(y,z) &= F_*(\phi(z) - \phi(y)) - (g_0(z) - g_0(y)) \\ &\leq \frac{c_1(y,z) + g_0(z) - g_(y)}{\phi(z) - \phi(y)} \cdot (\phi(z) - \phi(y)) - (g_0(z) - g_0(y)) = c_1(y,z). \end{aligned}$$

Then applying Itô's formula yields

$$\begin{aligned} f(X(t \wedge \sigma_n)) &= f(x_0) + \int_0^{t\wedge\sigma_n} Af(X(s))\, ds + \int_0^{t\wedge\sigma_n} f'(X(s))\, dW(s) \\ &\quad + \sum_{k=1}^{\infty} I_{\{\tau_k \leq t\wedge\sigma_n\}} Bf(X(\tau_k-), X(\tau_k)) \\ &\leq f(x_0) + \int_0^{t\wedge\sigma_n} [c_0(X(s)) - F_*]\, ds + \int_0^{t\wedge\sigma_n} f'(X(s))\, dW(s) \\ &\quad + \sum_{k=1}^{\infty} I_{\{\tau_k \leq t\wedge\sigma_n\}} c_1(X(\tau_k-), X(\tau_k)). \end{aligned}$$

Taking expectations on both sides, and rearranging the terms, it follows that

$$\begin{aligned} F_*\mathbb{E}_{x_0}[t \wedge \sigma_n] &\leq f(x_0) - \mathbb{E}_{x_0}[f(X(t \wedge \sigma_n))] + \mathbb{E}_{x_0}\left[\int_0^{t\wedge\sigma_n} c_0(X(s))ds\right] \\ &\quad + \mathbb{E}_{x_0}\left[\sum_{k=1}^{\infty} I_{\{\tau_k \leq t\wedge\sigma_n\}} c_1(X(\tau_k-), X(\tau_k))\right] \\ &\leq f(x_0) - K + \mathbb{E}_{x_0}\left[\int_0^{t\wedge\sigma_n} c_0(X(s))\, ds\right] \\ &\quad + \mathbb{E}_{x_0}\left[\sum_{k=0}^{\infty} c_1(X(\tau_k-), X(\tau_k)) I_{\{\tau_k \leq t\wedge\sigma_n\}}\right], \end{aligned}$$

where the last inequality follows from the observation that $f(x) = \frac{1}{6}\,x^4 - F_* x^2 \geq K > -\infty$ for some constant K. Letting $n \to \infty$, we know $\sigma_n \to \infty$ almost surely so the monotone convergence theorem yields

$$F_* t \leq f(x_0) - K + \mathbb{E}_{x_0}\left[\int_0^t c_0(X(s))\,ds + \sum_{k=0}^{\infty} c_1(X(\tau_k-), X(\tau_k))I_{\{\tau_k \leq t\}}\right].$$

Dividing by t and letting $t \to \infty$, we obtain

$$\begin{aligned} F_* &\leq \limsup_{t\to\infty} t^{-1}\mathbb{E}_{x_0}\left[\int_0^t c_0(X(s))\,ds + \sum_{k=1}^{\infty} c_1(X(\tau_k-), X(\tau_k))I_{\{\tau_k \leq t\}}\right] \\ &= J(\tau, Y). \end{aligned}$$

□

References

1. Bensoussan, A., Lions, J.-L.: Impulse control and quasi-variational inequalities. Bordas Editions (1984)
2. Chen, X., Sim, M., Simchi-Levi, D., Sun, P.: Risk aversion in inventory management. Oper. Res. **55**(5), 828–842 (2007)
3. Eastham, J.F., Hastings, K.J.: Optimal impulse control of portfolios. Math. Oper. Res. **13**(4), 588–605 (1988)
4. Harrison, J.M., Sellke, T.M., Taylor, A.J.: Impulse control of brownian motion. Math. Oper. Res. **8**(3), 454–466 (1983)
5. Helmes, K., Stockbridge, R.H.: Construction of the value function and optimal rules in optimal stopping of one-dimensional diffusions. Adv. Appl. Probab. **42**(1), 158–182 (2010)
6. Kurtz, T.G., Stockbridge, R.H.: Existence of Markov controls and characterization of optimal Markov controls. SIAM J. Control Optim. **36**(2), 609–653 (1998). (electronic)
7. 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
8. Lions, P.L., Perthame, B.: Quasi-variational inequalities and ergodic impulse control. SIAM J. Control Optim. **24**(4), 604–615 (1986)
9. Manne, A.S.: Linear programming and sequential decisions. Manag. Sci. **6**, 259–267 (1960)
10. Melas, D., Zervos, M.: An ergodic impulse control model with applications. In: Piunovskiy, A.B. (ed.) Modern Trends in Controlled Stochastic Processes, pp. 161–180. Luniver Press, Frome (2010)
11. Ormeci, M., Dai, J.G., Vate, J.V.: Impulse control of brownian motion: the constrained average cost case. Oper. Res. **56**(3), 618–629 (2008)
12. Piunovskiy, A.B.: Examples in Markov Decision Processes. Imperial College Press, London (2013)
13. Robin, M.: On some impulse control problems with long run average cost. SIAM J. Control Optim. **19**(3), 333–358 (1981)
14. Stockbridge, R.H.: Time-average control of martingale problems: existence of a stationary solution. Ann. Probab. **18**(1), 190–205 (1990)
15. Stockbridge, R.H.: Time-average control of martingale problems: a linear programming formulation. Ann. Probab. **18**(1), 206–217 (1990)

Impulse Control of Standard Brownian Motion: Discounted Criterion

Kurt Helmes[1], Richard H. Stockbridge[2(✉)], and Chao Zhu[2]

[1] Institut für Operations Research, Humboldt-Universität zu Berlin, Berlin, Germany
helmes@wiwi.hu-berlin.de
[2] Department of Mathematical Sciences, University of Wisconsin – Milwaukee, Milwaukee, WI 53201, USA
{stockbri,zhu}@uwm.edu

Abstract. This paper examines the impulse control of a standard Brownian motion under a discounted criterion. In contrast with the dynamic programming approach, this paper first imbeds the stochastic control problem into an infinite-dimensional linear program over a space of measures and derives a simpler nonlinear optimization problem that has a familiar interpretation. Optimal solutions are obtained for initial positions in a restricted range. Duality theory in linear programming is then used to establish optimality for arbitrary initial positions.

Keywords: Impulse control · Discounted criterion · Infinite dimensional linear programming · Expected occupation measures

1 Introduction

When one seeks to control a stochastic process and every intervention incurs a strictly positive cost, one must select a sequence of separate intervention times and amounts. The resulting stochastic problem is therefore an impulse control problem in which the decision maker seeks to either maximize a reward or minimize a cost. This paper continues the examination of the impulse control of Brownian motion. It considers a discounted cost criterion while a companion paper [5] studies the long-term average criterion. The aim of the paper is to illustrate a solution approach which first imbeds the stochastic control problem into an infinite-dimensional linear program over a space of measures and then reduces the linear program to a simpler nonlinear optimization. Contrasting with the long-term average paper, the dependence of the value function on the initial position of the process requires the use of duality in linear programming to obtain a complete solution.

Impulse control problems have been extensively studied using a quasi-variational approach; now classical works include [1,3] while the recent paper [2]

This research was supported in part by National Science Foundation under grant DMS-1108782 and by grant award 246271 from the Simons Foundation.

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 158–169, 2014.
DOI: 10.1007/978-3-662-45504-3_15

examines a Brownian inventory model. This paper extends a linear programming approach used on optimal stopping problems [4]. See [5] for additional references.

Let W be a standard Brownian motion process with natural filtration $\{\mathcal{F}_t\}$. An impulse control policy consists of a pair of sequences $(\tau, Y) := \{(\tau_k, Y_k) : k \in \mathbb{N}\}$ in which τ_k is the $\{\mathcal{F}_t\}$-stopping time of the kth impulse and the $\mathcal{F}_{\tau_k}$-measurable variable Y_k gives the kth impulse size. The sequence $\{\tau_k : k \in \mathbb{N}\}$ is required to be non-decreasing, a natural assumption in that intervention $k+1$ must occur no earlier than intervention k. For a policy (τ, Y), the impulse-controlled Brownian motion process is given by

$$X(t) = x_0 + W(t) + \sum_{k=1}^{\infty} I_{\{\tau_k \le t\}} Y_k.$$

The goal is to control the (discounted) second moment of X subject to (discounted) fixed and proportional costs for interventions. Let (τ, Y) be an impulse control policy. Define $c_0(x) = x^2$. Let $k_1 > 0$ denote the fixed costs incurred for each intervention and let $k_2 \ge 0$ be a cost proportional to the size of the intervention. Define the impulse cost function $c_1(y, z) = k_1 + k_2|z - y|$, in which y denotes the pre-jump location of X (typically far from 0) and z denotes the post-jump location of X which is thought to be close to 0. Let $\alpha > 0$ denote the discount rate. The objective function is

$$J(\tau, Y; x_0) = \mathbb{E}_{x_0}\left[\int_0^{\infty} e^{-\alpha s} c_0(X(s))\, ds + \sum_{k=1}^{\infty} I_{\{\tau_k < \infty\}} e^{-\alpha \tau_k} c_1(X(\tau_k-), X(\tau_k))\right]. \tag{1}$$

The controller must balance the desire to keep the process X near 0 so as to have a small second moment against the desire to limit the number and/or sizes of interventions so as to have a small impulse cost. Since the goal is to minimize the objective function, impulse control policies having $J(\tau, Y; x_0) = \infty$ are undesirable. We therefore restrict attention to the impulse policies for which $J(\tau, Y; x_0)$ is finite. Denote this class of *admissible* controls by $\mathcal{A}$.

We make five important observations about impulse policies. Firstly, "0-impulses" which do not change the state only increase the cost so can be excluded from consideration. Secondly, the symmetry of the dynamics and costs means that any impulse (τ_k, Y_k) which would cause $\text{sgn}(X(\tau_k)) = -sgn(X(\tau_k-))$ on a set of positive probability will have no greater cost (smaller cost when $k_2 > 0$) by replacing the impulse with one for which $\tilde{X}(\tau_k) = \text{sgn}(X(\tau_k-))|X(\tau_k)|$. Thus we can also restrict analysis to those policies for which all impulses keep the process on the same side of 0. Next, any policy (τ, Y) with $\lim_{k\to\infty} \tau_k =: \tau_\infty < \infty$ on a set of positive probability will have infinite cost so for every admissible policy $\tau_k \to \infty$ *a.s.* as $k \to \infty$. Next let (τ, Y) be a policy for which there is some k such that $\tau_k = \tau_{k+1}$ on a set of positive probability. Again due to the presence of the fixed intervention cost k_1, the total cost up to time τ_{k+1} will be at least $k_1 \mathbb{E}[e^{-\alpha\tau_k} I(\tau_k = \tau_{k+1})]$ smaller by combining these interventions into a

single intervention on this set. Hence we may restrict policies to those for which $\tau_k < \tau_{k+1}$ *a.s.* for each k.

The final observation is similar. Suppose (τ, Y) is a policy such that on a set G of positive probability $\tau_k < \infty$ and $|X(\tau_k)| > |X(\tau_k-)|$ for some k. Consider a modification of this impulse policy and resulting process $\tilde{X}$ which simply fails to implement this impulse on G. Define the stopping time $\sigma = \inf\{t > \tau_k : |X(t)| \leq |\tilde{X}(t)|\}$. Notice that the running costs accrued by $\tilde{X}$ over $[\tau_k, \sigma)$ are smaller than those accrued by X. Finally, at time σ, introduce an intervention on the set G which moves the $\tilde{X}$ process so that $\tilde{X}(\sigma) = X(\sigma)$. This intervention will incur a cost which is smaller than the cost for the process X at time τ_k. As a result, we may restrict the impulse control policies to those for which every impulse decreases the distance of the process from the origin.

2 Restricted Problem and Measure Formulation

The solution of the impulse control problem is obtained by first considering a subclass of the admissible impulse control pairs.

Condition 1. *Let $\mathcal{A}_1 \subset \mathcal{A}$ be those policies (τ, Y) such that the resulting process X is bounded; that is, for $(\tau, Y) \in \mathcal{A}_1$, there exists some $M < \infty$ such that $|X(t)| \leq M$ for all $t \geq 0$.*

Note that for each $M > 0$, any impulse control which has the process jump closer to 0 whenever $|X(t-)| = M$ is in the class $\mathcal{A}_1$ so this collection is non-empty. The bound is not required to be uniform for all $(\tau, Y) \in \mathcal{A}_1$. The restricted impulse control problem is one of minimizing $J(\tau, Y; x_0)$ over all policies $(\tau, Y) \in \mathcal{A}_1$.

We capture the expected behavior of the process and impulses with discounted measures. Let $(\tau, Y) \in \mathcal{A}_1$ be given and consider $f \in C^2(\mathbb{R})$. Then upon letting $t \to \infty$ after taking expectations, the general Dynkin's formula results in

$$\begin{aligned} f(x_0) = \mathbb{E}_{x_0}\left[\int_0^\infty e^{-\alpha s}[\alpha f(X(s)) - (1/2)f''(X(s))]\,ds\right] \\ + \mathbb{E}_{x_0}\left[\sum_{k=0}^{\infty} I_{\{\tau_k<\infty\}} e^{-\alpha\tau_k}[f(X(\tau_k-)) - f(X(\tau_k))]\right], \end{aligned} \tag{2}$$

in which the transversality condition $\lim_{t\to\infty} \mathbb{E}_{x_0}\left[e^{-\alpha t} f(X(t))\right] = 0$ follows from the boundedness of X. Note the generator of the Brownian motion process is $Af(x) = (1/2)f''(x)$. To simplify notation, define $Bf(y,z) = f(y) - f(z)$.

Define the discounted expected occupation measure μ_0 and the discounted impulse measure μ_1 such that for each $G, G_1, G_2 \subset \mathbb{R}$,

$$\begin{aligned} \mu_0(G) &= \mathbb{E}_{x_0}\left[\int_0^\infty e^{-\alpha s} I_G(X(s))\,ds\right] \\ \mu_1(G_1 \times G_2) &= \mathbb{E}_{x_0}\left[\sum_{k=0}^{\infty} I_{\{\tau_k<\infty\}} e^{-\alpha\tau_k} I_{G_1\times G_2}(X(\tau_k-), X(\tau_k))\right]. \end{aligned} \tag{3}$$

Notice that the total mass of μ_0 is $1/\alpha$ while μ_1 is a finite measure since $J(\tau, Y; x_0)$ is finite. Rewriting the objective function and Dynkin's formula in terms of these measures imbeds the impulse control problem in the linear program

$$\begin{cases} \text{Min.} & \displaystyle\int c_0 \, d\mu_0 \qquad + \int c_1 \, d\mu_1 \\ \text{S.t.} & \displaystyle\int (\alpha f - Af) \, d\mu_0 + \int Bf \, d\mu_1 = f(x_0), \quad \forall f \in C^2. \end{cases} \tag{4}$$

We now wish to introduce an auxiliary linear program derived from (4) which only has the μ_1 measure as its variable and has fewer constraints. Define $\phi(x) = e^{-\sqrt{2\alpha}x}$ and $\psi(x) = e^{\sqrt{2\alpha}x}$. Notice that ϕ is a strictly decreasing solution while ψ is a strictly increasing solution of the homogeneous equation $\alpha f - Af = 0$. For each $(\tau, Y) \in \mathcal{A}_1$, the resulting process X is bounded so we can use both ϕ and ψ in (2). This results in the two constraints

$$\int B\phi(y,z) \, \mu_1(dy \times dz) = \phi(x_0) \quad \text{and} \quad \int B\psi(y,z) \, \mu_1(dy \times dz) = \psi(x_0) \tag{5}$$

which only constrain the measure μ_1. Note that the monotonicity and positivity of both ϕ and ψ require the support of μ_1 to be such that the two integrals in (5) are positive. We can also take advantage of the symmetry inherent in the problem. Define $p_0(x) = \cosh(\sqrt{2\alpha}x)$. Then averaging the two constraints (5) yields

$$\int \frac{Bp_0(y,z)}{p_0(x_0)} \, \mu_1(dy \times dz) = 1.$$

Using $g_0(x) = (\alpha x^2 + 1)/\alpha^2$ in (2), where again the boundedness of X implies that the transversality condition is satisfied, yields

$$\begin{aligned} &\mathbb{E}_{x_0}\left[\int_0^\infty e^{-\alpha s} c_0(X(s)) \, ds\right] \\ &= \frac{\alpha x_0^2 + 1}{\alpha^2} - \mathbb{E}_{x_0}\left[\sum_{k=0}^{\infty} I_{\{\tau_k < \infty\}} e^{-\alpha \tau_k} Bg_0(X(\tau_k-), X(\tau_k))\right]. \end{aligned} \tag{6}$$

Let $[c_1 - Bg_0]$ denote the sum of the two functions c_1 and Bg_0. Using (6) in (1) establishes that

$$J(\tau, Y; x_0) = \frac{\alpha \, x_0^2 + 1}{\alpha^2} + \mathbb{E}_{x_0}\left[\sum_{k=0}^{\infty} I_{\{\tau_k < \infty\}} e^{-\alpha \tau_k} [c_1 - Bg_0](X(\tau_k-), X(\tau_k))\right]$$

and hence that

$$J(\tau, Y; x_0) = \frac{\alpha \, x_0^2 + 1}{\alpha^2} + \int [c_1 - Bg_0](y,z) \, \mu_1(dy \times dz) \tag{7}$$

so the objective function value only depends on the measure μ_1. Since the objective function for each $(\tau, Y) \in \mathcal{A}_1$ has the affine term $g_0(x_0)$, it may be ignored for the purposes of optimization but it must be included to obtain the correct value for the objective function. Now form the auxiliary linear program

$$\begin{cases} \text{Min.} & \displaystyle\int [c_1 - Bg_0](y,z)\,\mu_1(dy \times dz) \\ \text{S.t.} & \displaystyle\int \frac{Bp_0(y,z)}{p_0(x_0)}\,\mu_1(dy \times dz) = 1. \end{cases} \tag{8}$$

Let $V_1(x_0)$ denote the value of the impulse control problem over policies in $\mathcal{A}_1$, V_{lp} denote the value of (4) and V_{aux} denote the value of (8). The following proposition is immediate.

Proposition 2. $V_{aux}(x_0) \le V_{lp}(x_0) \le V_1(x_0)$.

Remark 3. *Our analysis will also involve other auxiliary linear programs as well. One will replace the single constraint in* (8) *with the pair of constraints* (5) *while another will limit the constraints in* (4) *to a single function. Each auxiliary program will provide a lower bound on* $V_{lp}(x_0)$ *and hence on* $V_1(x_0)$.

2.1 Nonlinear Optimization and Partial Solution

Recall, the admissible impulse policies can be (and are) limited to those for which impulses move X closer to the origin. As a result, the integrand $Bp_0 > 0$ and the constraint of (8) implies that the feasible measures μ_1 of (8) are those for which $Bp_0/p_0(x_0)$ is a probability density. For a feasible μ_1, let $\tilde{\mu}_1$ be the probability measure $\frac{Bp_0}{p_0(x_0)}\,\mu_1$. Thus we can write the objective function as

$$\int [c_1 - Bg_0]\,d\mu_1 = \left(\int \frac{c_1 - Bg_0}{Bp_0}\,d\tilde{\mu}_1 \right) p_0(x_0).$$

Since the goal is to minimize the cost, a lower bound is given by the minimal value of F scaled by the constant $p_0(x_0)$, where

$$F(y,z) := \frac{c_1(y,z) - Bg_0(y,z)}{Bp_0(y,z)}.$$

Moreover, should the infimum be attained at some pair (y_*, z_*), then the probability measure $\tilde{\mu}_1(\cdot)$ putting unit point mass on (y_*, z_*) would achieve the lower bound and identify an optimal μ_1 measure for the auxiliary linear program. To solve the stochastic problem, one would need to connect the measure μ_1 back to an admissible impulse control policy in the class $\mathcal{A}_1$ in such a way that the resulting μ_1 measure would be given by (3).

Remark 4. *The objective function* $p_0(x_0)F$ *has a natural interpretation. First observe that* $Bp_0(y,z) = \cosh(\sqrt{2\alpha}y) - \cosh(\sqrt{2\alpha}z)$ *so*

$$p_0(x_0)F(y,z) = [c_1(y,z) - Bg_0(y,z)] \cdot \frac{\cosh(\sqrt{2\alpha}x_0)}{\cosh(\sqrt{2\alpha}y)} \cdot \sum_{n=0}^{\infty} \left(\frac{\cosh(\sqrt{2\alpha}z)}{\cosh(\sqrt{2\alpha}y)} \right)^n.$$

It can be shown that the first fraction gives the expected discount for the time it takes X to reach $\{\pm y\}$ when starting at x_0. The ratio $\frac{\cosh(\sqrt{2\alpha}z)}{\cosh(\sqrt{2\alpha}y)}$ then gives the expected discount for the time it takes X to again reach $\{\pm y\}$ but this time starting at $\pm z$ so the sum represents the expected discounting for infinitely many cycles. By symmetry, the initial term gives the cost for impulsing from $\pm y$ to $\pm z$ along with the second moment. The minimization therefore optimizes the expected cost over a particular class of impulse policies. We emphasize that the linear program imbedding is not *restricted to these policies.*

Proposition 5. *There exists pairs (y_*, z_*) and $(-y_*, -z_*)$ such that*

$$F(y_*, z_*) = F(-y_*, -z_*) = \inf_{(y.z):|z|\le|y|} F(y, z). \tag{9}$$

Moreover, the minimizing pair (y_*, z_*) having nonnegative components is unique.

Proof. First observe

$$F(y, z) = \frac{k_1 + k_2|y - z| + (z^2 - y^2)/\alpha}{\cosh(\sqrt{2\alpha}y) - \cosh(\sqrt{2\alpha}z)}$$

so there exists some pairs (y, z) for which $F(y, z) < 0$ since the difference of the quadratic terms is negative and will dominate the constant and linear terms in the numerator. A straightforward asymptotic analysis show that $F(y, z)$ is asymptotically nonnegative when $y \to \infty$, $z \to \infty$ or $|y - z| \to 0$. Therefore F achieves its minimum at some point (y_*, z_*).

Notice that F is symmetric about 0 in that $F(-y, -z) = F(y, z)$ so it is sufficient to analyze F on the domain $0 \le z \le y$. The first-order optimality conditions on F are

$$\begin{aligned}
0 = \frac{\partial F}{\partial y}(y_*, z_*) &= \frac{(k_2 - 2y_*/\alpha)[\cosh(\sqrt{2\alpha}\, y_*) - \cosh(\sqrt{2\alpha}\, z_*)]}{[\cosh(\sqrt{2\alpha}\, y_*) - \cosh(\sqrt{2\alpha}\, z_*)]^2} \\
&\quad - \frac{\sqrt{2\alpha}\,[k_1 + k_2(y_* - z_*) + (z_*^2 - y_*^2)/\alpha]\sinh(\sqrt{2\alpha}\, y_*)}{[\cosh(\sqrt{2\alpha}\, y_*) - \cosh(\sqrt{2\alpha}\, z_*)]^2}, \\
0 = \frac{\partial F}{\partial z}(y_*, z_*) &= \frac{(-k_2 + 2z_*/\alpha)[\cosh(\sqrt{2\alpha}\, y_*) - \cosh(\sqrt{2\alpha}\, z_*)]}{[\cosh(\sqrt{2\alpha}\, y_*) - \cosh(\sqrt{2\alpha}\, z_*)]^2} \\
&\quad + \frac{\sqrt{2\alpha}\,[k_1 + k_2(y_* - z_*) + (z_*^2 - y_*^2)/\alpha]\sinh(\sqrt{2\alpha}\, z_*)}{[\cosh(\sqrt{2\alpha}\, y_*) - \cosh(\sqrt{2\alpha}\, z_*)]^2}.
\end{aligned}$$

The minimizing pair (y_*, z_*) will be interior to the region since $\frac{\partial F}{\partial z}(y_*, 0) = \frac{-k_2}{\cosh(\sqrt{2\alpha}\, y_*) - 1} < 0$.

Simple algebra now leads to the following systems of nonlinear equations for (y_*, z_*):

$$\begin{aligned} k_2\alpha - 2y_* &= \alpha\sqrt{2\alpha}\,\sinh(\sqrt{2\alpha}\,y_*)\cdot F(y_*,z_*),\\ k_2\alpha - 2z_* &= \alpha\sqrt{2\alpha}\,\sinh(\sqrt{2\alpha}\,z_*)\cdot F(y_*,z_*). \end{aligned} \tag{10}$$

The fact that the minimal value of F is negative implies $y_* > z_* > k_2\alpha/2$. Solving for $-\alpha\sqrt{2\alpha}\,F(y_*,z_*)$ in each equation shows that at an optimal pair (y_*,z_*),

$$-\alpha\sqrt{2\alpha}\,F(y_*,z_*) = \frac{2z_* - k_2\alpha}{\sinh(\sqrt{2\alpha}\,z_*)} = \frac{2y_* - k_2\alpha}{\sinh(\sqrt{2\alpha}\,y_*)}\,.$$

A straightforward analysis of the function $h(x) = [2x - k_2\alpha]/\sinh(\sqrt{2\alpha}\,x)$ on the domain $[k_2\alpha/2,\infty)$ shows that the level sets of h consist of two-point sets and so on the region $0 \le z \le y$, the pair (y_*,z_*) is unique. □

Now that the lower bound given in (9) is determined, it is important to connect an optimizing μ_1^* with an admissible impulse control policy $(\tau,Y) \in \mathcal{A}_1$. The existence of two minimizing pairs (y_*,z_*) and $(-y_*,-z_*)$ allows many auxiliary-LP-feasible measures μ_1 to place point masses at these two points and still achieve the lower bound. This observation leads to a solution to the restricted stochastic impulse control problem.

Theorem 6. *Let (y_*,z_*) be the pair having positive components that minimizes F as identified in Proposition 5. Consider initial positions $-y_* \le x_0 \le y_*$. Define the impulse control policy (τ^*,Y^*) as follows:*

$$\tau_1^* = \inf\{t \ge 0 : X(t-) = \pm y_*\} \quad \textit{and} \quad Y_1^* = sgn(X(\tau_1^*-))\cdot z_* - X(\tau_1^*-)$$

and for $k = 2,3,4,\ldots$, define

$$\tau_k^* = \inf\{t > \tau_{k-1} : X(t-) = \pm y_*\} \quad \textit{and} \quad Y_k^* = sgn(X(\tau_k^*-))\cdot z_* - X(\tau_k^*-).$$

Then (τ^*,Y^*) is an optimal impulse control pair for the restricted stochastic impulse control problem and the corresponding optimal value is

$$V_1(x_0) = \frac{\alpha x_0^2 + 1}{\alpha^2} + F(y_*,z_*)\cdot\cosh(\sqrt{2\alpha}\,x_0). \tag{11}$$

Proof. The measure μ_1^* defined from (τ^*,Y^*) using (3) is concentrated on the two points $(-y_*,-z_*)$ and (y_*,z_*). Since the process resulting from the admissible impulse control pair (τ^*,Y^*) remains bounded, conditions (5) can be used to obtain the masses:

$$\begin{aligned} &\mu_1^*(-y_*,-z_*)\\ &\quad = \frac{\phi(x_0)[\psi(y_*)-\psi(z_*)] - \psi(x_0)[\phi(y_*)-\phi(z_*)]}{[\phi(-y_*)-\phi(-z_*)][\psi(y_*)-\psi(z_*)] - [\psi(-y_*)-\psi(-z_*)][\phi(y_*)-\phi(z_*)]}\,,\\ &\mu_1^*(y_*,z_*)\\ &\quad = \frac{\psi(x_0)[\phi(-y_*)-\phi(-z_*)] - \phi(x_0)[\psi(-y_*)-\psi(-z_*)]}{[\phi(-y_*)-\phi(-z_*)][\psi(y_*)-\psi(z_*)] - [\psi(-y_*)-\psi(-z_*)][\phi(y_*)-\phi(z_*)]}\,. \end{aligned}$$

Recall $\phi(x) = e^{-\sqrt{2\alpha}\,x}$ and $\psi(x) = e^{\sqrt{2\alpha}\,x}$ so $\phi(-x) = \psi(x)$ and $\psi(-x) = \phi(x)$. As a result these expressions simplify to

$$\mu_1^*(-y_*, -z_*) = \frac{\phi(x_0)[\psi(y_*) - \psi(z_*)] - \psi(x_0)[\phi(y_*) - \phi(z_*)]}{[\psi(y_*) - \psi(z_*)]^2 - [\phi(y_*) - \phi(z_*)]^2},$$
$$\mu_1^*(y_*, z_*) = \frac{\psi(x_0)[\psi(y_*) - \psi(z_*)] - \phi(x_0)[\phi(y_*) - \phi(z_*)]}{[\psi(y_*) - \psi(z_*)]^2 - [\phi(y_*) - \phi(z_*)]^2}.$$

It is now straightforward to verify that $J(\tau^*, Y^*; x_0)$ equals the value in (11).□

2.2 Full Solution

Theorem 6 solves the problem for initial positions x_0 with $|x_0| \le y_*$. The issue is now one of determining the optimal value and an optimal impulse control pair when $|x_0| > y_*$. From an intuitive point of view, $|x_0| < y_*$ has an optimal control which waits until the state process first hits $\pm y_*$ before having an impulse so one might expect an impulse to occur immediately when $|x_0| \ge y_*$. Since two impulses at the same instant are no better than one, one would anticipate that the after-jump location might be $z \in (-y_*, y_*)$. The cost of an immediate jump from x_0 to z followed by using an optimal impulse control is

$$\begin{aligned} g(z) &:= \frac{\alpha x_0^2 + 1}{\alpha^2} + k_1 + k_2(x_0 - z) + \frac{z^2 - x_0^2}{\alpha} + V_1(z) \\ &= \frac{\alpha z^2 + 1}{\alpha^2} + k_1 + k_2(x_0 - z) + V_1(z). \end{aligned}$$

Solving $g'(z) = 0$ to find a minimizer results in

$$0 = -k_2 + 2z/\alpha + \sqrt{2\alpha}\,F(y_*, z_*)\sinh(\sqrt{2\alpha}\,z),$$

which is the first order condition (10) for which both y_* and z_* are solutions. An impulse to y_* would be followed by an immediate jump to z_* and incur two fixed costs whereas a single jump directly to z_* would cost less. This line of reasoning indicates that a single jump to z_* could be an optimal initial impulse.

The goal is to verify that this intuitive reasoning is correct. Define

$$\widehat{V}(y) = \begin{cases} k_1 + k_2(|y| - z_*) + V_1(-z_*), & y \le -y_*, \\ V_1(y), & -y_* \le y \le y_*, \\ k_1 + k_2(y - z_*) + V_1(z_*), & y \ge y_*. \end{cases}$$

For $|y| > y_*$, the function $\widehat{V}$ is the cost associated with the process starting at initial position y, having an instantaneous jump from y to $\operatorname{sgn}(y)\,z_*$ and then using the optimal impulse control policy of Theorem 6 thereafter. The following lemma is fairly straightforward so its proof is left to the reader.

Lemma 7. $\widehat{V} \in C^1(\mathbb{R}) \cap C^2(\mathbb{R}\backslash\{\pm y_*\})$.

The function $\widehat{V}$ therefore has sufficient regularity to use in (2). We now consider the new auxiliary linear program

$$\begin{cases} \text{Min.} & \displaystyle\int_{\mathbb{R}\backslash\{\pm y_*\}} c_0(x)\,\mu_0(dx) \quad + \quad \int_{\mathbb{R}^2} c_1(y,z)\,\mu_1(dy\times dz) \\ \text{S.t.} & \displaystyle\int_{\mathbb{R}\backslash\{\pm y_*\}} [\alpha\widehat{V}(x) - A\widehat{V}(x)]\,\mu_0(dx) + \int_{\mathbb{R}^2} [\widehat{V}(y) - \widehat{V}(z)]\,\mu_1(dy\times dz) \\ & \qquad = \widehat{V}(x_0) \end{cases} \tag{12}$$

and its dual (having sole variable w)

$$\begin{cases} \text{Max.} & \widehat{V}(x_0)\cdot w \\ \text{S.t.} & \left(\alpha\widehat{V}(x) - A\widehat{V}(x)\right)\cdot w \le\ c_0(x), \quad x \ne \pm y_*, \\ & \left(\widehat{V}(y) - \widehat{V}(z)\right)\cdot w \ \le c_1(y,z), \quad \forall y,z\in\mathbb{R}. \end{cases} \tag{13}$$

Observe that each linear program has feasible points with costs that are finite. A straightforward weak duality argument therefore shows that each value of (13) corresponding to a feasible variable w is no greater than any value of (12) for a feasible pair of measures and hence the value of (13) is a lower bound on the value of the restricted impulse control problem. Since $\widehat{V}(x_0) > 0$, one seeks as large a positive value as possible for w.

Theorem 8. *The optimal value of* (13) *is* $\widehat{V}(x_0)$ *which is achieved when* $w_* = 1$.

Proof. By symmetry, it is sufficient to examine $x, y, z \ge 0$. Notice that for $0 \le x < y_*$, $\alpha\widehat{V}(x) - A\widehat{V}(x) = x^2 = c_0(x) \ge 0$ and hence the dual variable w cannot exceed 1. The question is whether $w = 1$ is feasible for (13) so examine the rest of the constraints with $w = 1$.

For $x > y_*$, $A\widehat{V}(x) = 0$ so the first constraint of (13) requires

$$0 \le x^2 - \alpha\left(k_1 + k_2(x - z_*) + V_1(z_*)\right) = x^2 - \alpha V_1(y_*).$$

Since the right-hand expression is an increasing function for $x \in [k_2\alpha/2, \infty)$, it suffices to verify its nonnegativity with $x = y_*$:

$$\begin{aligned} 0 \le y_*^2 - \alpha V_1(y_*) &= y_*^2 - \alpha\left(\frac{\alpha y_*^2 + 1}{\alpha^2} + F(y_*, z_*)\cosh(\sqrt{2\alpha}\,y_*)\right) \\ &= -\frac{1}{\alpha} + \frac{[2y_* - k_2\alpha]\cosh(\sqrt{2\alpha}\,y_*)}{\sqrt{2\alpha}\,\sinh(\sqrt{2\alpha}\,y_*)} \end{aligned}$$

in which (10) is used to obtain the last expression. This inequality can be rewritten as

$$\frac{\tanh(\sqrt{2\alpha}\,y_*)}{\sqrt{2\alpha}} \le y_* - k_2\alpha/2. \tag{14}$$

Since (y_*, z_*) is a minimizing pair of the function F, (14) holds and the first family of constraints of (13) is satisfied with $w = 1$.

Consider now the second family of constraints with $w = 1$. There are several cases to examine. When $0 \le y \le z$, monotonicity of $\widehat{V}$ on this range shows the condition is trivially satisfied. Next, for $0 \le z \le y \le y_*$, the constraint can be rewritten as

$$V_1(y) \le k_1 + k_2(y - z) + V_1(z).$$

The right-hand expression gives the cost of an immediate jump from y to z followed by an optimal impulse control policy thereafter whereas the left-hand side gives the optimal cost. Hence this inequality is satisfied. Now consider $y_* \le z < y$ and observe that $\widehat{V}(y) - \widehat{V}(z) = k_2(y - z) < k_1 + k_2(y - z)$. Finally, for $0 \le z < y_* < y$ and again using the definition of $\widehat{V}$, the second set of constraints in (13) is equivalent to

$$k_1 + k_2(y - z_*) + V_1(z_*) \le k_1 + k_2(y - z) + V_1(z)$$

or equivalently

$$\begin{aligned} k_2(y - y_*) + V_1(y_*) &= k_2(y - y_*) + [k_1 + k_2(y_* - z_*) + V_1(z_*)] \\ &\le k_2(y - y_*) + [k_1 + k_2(y_* - z) + V_1(z)]. \end{aligned}$$

This last inequality is true by the optimality of both the pair (y_*, z_*) and the function V_1 on $[-y_*, y_*]$ since the bracketed quantity on the right-hand side gives the cost associated with an initial impulse to z from y_* along with optimal impulse control policy starting from z. Thus the second family of constraints in (13) hold when $w = 1$. □

We now have the following result.

Theorem 9. *Let (y_*, z_*) be the optimizing pair for F having positive components. Define the impulse control policy (τ^*, Y^*) as follows;*

$$\tau_1^* = \inf\{t \ge 0 : |X(t-)| \ge y_*\} \quad \textit{and} \quad Y_1^* = \operatorname{sgn}(X(\tau_1^*-)) \cdot z_* - X(\tau_1^*-)$$

and for $k = 2, 3, 4, \ldots$, define

$$\tau_k^* = \inf\{t > \tau_{k-1} : X(t-) = \pm y_*\} \quad \textit{and} \quad Y_k^* = \operatorname{sgn}(X(\tau_k^*-)) \cdot z_* - X(\tau_k^*-).$$

Then (τ^*, Y^*) is an optimal impulse control pair for the restricted stochastic impulse control problem and the corresponding optimal value is $\widehat{V}(x_0)$.

Proof. The particular choice of (τ^*, Y^*) implies $\widehat{V}(x_0) \le V_{lp}(x_0) \le V_1(x_0) \le J(\tau^*, Y^*) = \widehat{V}(x_0)$. □

2.3 Solution for General Admissible Impulse Controls

The solution of Sect. 2.2 is restricted to those impulse control policies under which the process X remains bounded. It is necessary to show that no lower cost can be obtained by any policy which allows the process to be unbounded.

Theorem 10. *The impulse control policy (τ^*, Y^*) of Theorem 9 is optimal in the class of all admissible policies and $\widehat{V}(x_0)$ is the optimal value.*

Proof. This argument establishes that $\widehat{V}(x_0)$ is a lower bound on $J(\tau, Y; x_0)$ for every admissible impulse control policy. Theorem 9 then gives the existence of an optimal policy whose cost equals the lower bound.

Choose $(\tau, Y) \in \mathcal{A}$ and let X be the resulting controlled process. Suppose there exists some $K > 0$ such that $\liminf_{t\to\infty} \mathbb{E}_{x_0}[e^{-\alpha t}\widehat{V}(X(t))] \geq K$. Note that

$$\liminf_{t\to\infty} \mathbb{E}_{x_0}\left[e^{-\alpha t}\widehat{V}(X(t))\right] = \liminf_{t\to\infty} \mathbb{E}_{x_0}\left[e^{-\alpha t}\widehat{V}(X(t))I_{\{|X(t)|\geq y_*\}}\right]$$

so the linearity of $\widehat{V}$ on $\{x : |x| \geq y_*\}$ implies that $\mathbb{E}_{x_0}[|X(t)|I_{\{|X(t)|\geq y_*\}}]$ is asymptotically bounded below by $Ke^{\alpha t}$ as $t \to \infty$. Hence by Jensen's inequality for $\epsilon > 0$ and t large,

$$\mathbb{E}_{x_0}\left[X^2(t)\right] \geq \left(\mathbb{E}_{x_0}[\,|X(t)|I_{\{|X(t)|\geq y_*\}}]\right)^2 \geq K^2e^{2\alpha t} - \epsilon.$$

Using this estimate in (1) shows $J(\tau, Y; x_0) = \infty$.

Now suppose $J(\tau, Y; x_0) < \infty$ so $\liminf_{t\to\infty} \mathbb{E}_{x_0}[e^{-\alpha t}\widehat{V}(X(t))] = 0$. Then there exists a sequence $\{t_j : j \in \mathbb{N}\}$ such that $\lim_{j\to\infty} \mathbb{E}_{x_0}[e^{-\alpha t_j}\widehat{V}(X(t_j))] = 0$. Note that $|\widehat{V}'| \leq k_2$ so $\int_0^t e^{-\alpha s}\widehat{V}'(X(s))\,dW(s)$, $t \geq 0$, is a martingale. Thus the dual constraints, in conjunction with the finiteness of the expected cost, implies that Dynkin's formula holds when $t = t_j$ for each j. Hence

$$\begin{aligned}
\widehat{V}(x_0) &= \mathbb{E}_{x_0}\left[\int_0^{t_j} e^{-\alpha s}[\alpha\,\widehat{V}(X(s)) - A\widehat{V}(X(s))]\,ds\right] - E_{x_0}\left[e^{-\alpha t_j}\widehat{V}(X(t_j))\right] \\
&\quad + \mathbb{E}_{x_0}\left[\sum_{k=0}^{\infty} I_{\{\tau_k\leq t_j\}}e^{-\alpha\tau_k}B\widehat{V}(X(\tau_k-), X(\tau_k))\right] \\
&\leq \mathbb{E}_{x_0}\left[\int_0^{t_j} e^{-\alpha s}c_0(X(s))\,ds + \sum_{k=0}^{\infty} I_{\{\tau_k\leq t_k\}}e^{-\alpha\tau_k}c_1(X(\tau_k-), X(\tau_k))\right] \\
&\quad - \mathbb{E}_{x_0}\left[e^{-\alpha t_j}\widehat{V}(X(t_j))\right]
\end{aligned}$$

Letting $j \to \infty$, an application of the monotone convergence theorem on the first expectation and the convergence to 0 of second expectation establishes that $\widehat{V}(x_0)$ is a lower bound on the expected cost $J(\tau, Y; x_0)$. □

References

1. Bensoussan, A., Lions, J.-L.: Impulse Control and Quasi-Variational Inequalities. Bordas Editions. Gauthier-Villars, Paris (1984)
2. Dai, J.G., Yao, D.: Optimal control of Brownian inventory models with convex inventory cost, Part 2: Discount-optimal controls. Stoch. Syst. **3**(2), 500–573 (2013)
3. Harrison, J.M., Sellke, T.M., Taylor, A.J.: Impulse control of Brownian motion. Math. Oper. Res. **8**(3), 454–466 (1983)
4. Helmes, K., Stockbridge, R.H.: Construction of the value function and optimal rules in optimal stopping of one-dimensional diffusions. Adv. Appl. Probab. **42**(1), 158–182 (2010)
5. Helmes, K., Stockbridge, R.H., Zhu, C.: Impulse control of standard Brownian motion: long-term average criterion. In: Helmes, K., Stockbridge, R.H., Zhu, C. (eds.) System Modeling and Optimization, vol. 443, pp. 148–157 (2014)

On Target Control Synthesis Under Set-Membership Uncertainties Using Polyhedral Techniques

Elena K. Kostousova(✉)

N.N. Krasovskii Institute of Mathematics and Mechanics of the Ural Branch of the Russian Academy of Sciences, 16, S.Kovalevskaja Street, Ekaterinburg 620990, Russia
kek@imm.uran.ru

Abstract. Problems of feedback terminal target control for linear and bilinear uncertain systems are considered. We continue the development of control synthesis using polyhedral (parallelotope-valued) solvability tubes. The paper deals with two types of problems, where controls appear either additively or in the system matrix. For both problems, the cases without uncertainties, with additive parallelotope-bounded uncertainties, and also with interval uncertainties in coefficients of the system (a bilinear uncertainty) are considered. Ordinary differential equations, which describe the mentioned polyhedral solvability tubes, are presented for each of these cases. New control strategies, which can be calculated by explicit formulas on the base of the mentioned tubes, are proposed. Results of computer simulations are presented.

Keywords: Differential systems · Uncertain systems · Control synthesis · Polyhedral estimates · Parallelotopes · Interval analysis

1 Introduction

Problems of feedback terminal target control for linear and bilinear differential uncertain systems are considered. There are known approaches for solving problems like these, in particular, based on constructing solvability tubes and the extremal aiming strategies of N.N. Krasovskii [13,17]. The problem statement for linear systems, approaches for solving, and the tight interconnections between solvability tubes, the Pontriagin alternated integral, Hamilton-Jacobi-Bellman equations, and funnel equations can be found, for example, in [15–17].

Since practical construction of the mentioned tubes can be cumbersome, different numerical methods are devised, in particular, methods for approximating the set-valued integrals and for numerical solving the mentioned equations,

The research was partially supported by the Program of Basic Research of the Ural Branch of RAS (Project 12-P-1-1019), by the Russian Foundation for Basic Research (Grant 12-01-00043), and by the Program for the State Support of Leading Scientific Schools of Russian Federation (Grant 2692.2014.1).

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 170–180, 2014.
DOI: 10.1007/978-3-662-45504-3_16

including methods based on approximations of sets by arbitrary polytopes with a large number of vertices [2,4,21,22] (here and below we mention, as examples, only some references from numerous publications; see also references therein). Such methods are devised to obtain approximations as accurate as possible. But they can require much calculations, especially for large dimensional systems. Other techniques are based on estimates of sets by domains of some fixed shape such as ellipsoids and parallelepipeds, including boxes aligned with coordinate axes as in interval analysis [4,5,7–12,14,15,17,18,20]. The main advantage of such techniques is that they enable to obtain approximate/particular solutions using relatively simple tools (up to explicit formulas). More accurate approximations may be obtained by using the whole families (varieties) of such simple estimates (as was proposed by A.B. Kurzhanski) [8,12,15,17,18].

For linear differential systems, the constructive computation schemes for solving the feedback target control problems by means of ellipsoidal techniques were proposed [15,17] and then expanded to a polyhedral technique [8]. Here we continue the development of the polyhedral control synthesis using polyhedral (parallelotope-valued) solvability tubes. The paper deals with two types of problems, where the controls appear either additively or in the system matrix. For both problems, the cases without uncertainties, with additive uncertainties, and also with interval uncertainties in coefficients of the system (the bilinear uncertainty) are considered. Ordinary differential equations (ODE) for the mentioned polyhedral solvability tubes are presented. New control strategies, which can be calculated by explicit formulas on the base of the mentioned tubes, are proposed. In opposite to [8,15,17], they are concretized by explicit formulas when the state belongs to a tube. Also the polyhedral control synthesis for discrete-time systems is considered. The results of computer simulations are presented.

Note that there are also some works devoted to other approaches for solving different control problems under uncertainty and works concerning systems with bilinear uncertainties (see, for example, [1,4,6,19,20]).

The following notation is used below: $\mathbb{R}^n$ is the n-dimensional vector space; $\top$ is the transposition symbol; $\|x\|_2 = (x^\top x)^{1/2}$, $\|x\|_\infty = \max_{1\le i\le n} |x_i|$ are vector norms for $x = (x_1, x_2, \dots, x_n)^\top \in \mathbb{R}^n$; $\mathrm{e}^i = (0, \dots, 0, 1, 0, \dots, 0)^\top$ is the unit vector oriented along the axis $0x_i$ (the unit stands at position i); $\mathrm{e} = (1, 1, \dots, 1)^\top$; $\mathbb{R}^{n\times m}$ is the space of real $n \times m$-matrices $A = \{a_i^j\} = \{a^j\}$ (with columns a^j); I is the identity matrix; 0 is the zero matrix (vector); $\operatorname{Abs} A = \{|a_i^j|\}$ for $A = \{a_i^j\}$; $\operatorname{diag}\pi$, $\operatorname{diag}\{\pi_i\}$ are the diagonal matrix A with $a_i^i = \pi_i$ (π_i are the components of the vector π); $\det A$ is the determinant of A; $\operatorname{tr} A = \sum_{i=1}^n a_i^i$ is the trace of A; $\|A\| = \max_{1\le i\le n} \sum_{j=1}^m |a_i^j|$ for $A \in \mathbb{R}^{n\times m}$; $\operatorname{int}\mathcal{X}$ is the set of interior points of the set $\mathcal{X} \subset \mathbb{R}^n$; the notation of the type $k = 1, \dots, N$ is used instead of $k = 1, 2, \dots, N$.

2 Problems Formulation

Consider the controlled system with a given terminal set $\mathcal{M}$ ($x \in \mathbb{R}^n$ is the state):

$$\dot{x} = (A(t) + U(t) + V(t))\, x + u(t) + v(t), \quad t \in T = [0, \theta]. \tag{1}$$

Here $A(t) \in \mathbb{R}^{n\times n}$ is a given matrix function; Lebesgue measurable functions $U(t) \in \mathbb{R}^{n\times n}$ and $u(t) \in \mathbb{R}^n$ serve as controls and satisfy either (2) or (3):

$$U(t) \equiv 0, \quad u(t) \in \mathcal{R}(t), \qquad \text{a.e. } t \in T, \tag{2}$$

$$U(t) \in \mathcal{U}(t) = \{U \in \mathbb{R}^{n\times n} | \operatorname{Abs}(U - \tilde{U}(t)) \le \hat{U}(t)\}, \quad u(t) \equiv 0, \quad \text{a.e. } t \in T; \tag{3}$$

$V(t) \in \mathbb{R}^{n\times n}$ and $v(t) \in \mathbb{R}^n$ stand for unknown disturbances and satisfy

$$V(t) \in \mathcal{V}(t) = \{V \in \mathbb{R}^{n\times n} | \operatorname{Abs}(V - \tilde{V}(t)) \le \hat{V}(t)\}, \quad v(t) \in \mathcal{Q}(t), \quad \text{a.e. } t \in T. \tag{4}$$

Matrix and vector inequalities $(\le, <, \ge, >)$ here and below are understood componentwise. We presume the sets $\mathcal{R}(t)$, $\mathcal{Q}(t)$, and $\mathcal{M}$ to be parallelotopes and a parallelepiped respectively:

$$\mathcal{R}(t) = \mathcal{P}[r(t), \bar{R}(t)], \ \bar{R}(t) \in \mathbb{R}^{n\times n_1}, \quad \mathcal{Q}(t) = \mathcal{P}[q(t), \bar{Q}(t)], \ \bar{Q}(t) \in \mathbb{R}^{n\times n_2},$$
$$\mathcal{M} = \mathcal{P}(p_{\mathrm{f}}, P_{\mathrm{f}}, \pi_{\mathrm{f}}) = \mathcal{P}[p_{\mathrm{f}}, \bar{P}_{\mathrm{f}}], \qquad \bar{P}_{\mathrm{f}} \in \mathbb{R}^{n\times n}, \ \det \bar{P}_{\mathrm{f}} \neq 0; \tag{5}$$

$r(t)$, $\bar{R}(t)$, $q(t)$, $\bar{Q}(t)$, as well as $A(t)$, $\tilde{U}(t)$, $\hat{U}(t) \ge 0$, $\tilde{V}(t)$, $\hat{V}(t) \ge 0$, are known continuous vector and matrix functions; the parallelepiped $\mathcal{M}$ is nondegenerate.

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 $\det P \neq 0$, $\|p^i\|_2 = 1$[1]; $\pi \in \mathbb{R}^n$, $\pi \ge 0$. It may be said that p determines the center of the parallelepiped, P is the orientation matrix, p^i are the "directions" and π_i are 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 $\det \bar{P} \neq 0$.

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 $\mathrm{e} = (1, 1, \ldots, 1)^\top$.

We can consider the above system for the following cases: (I) *without uncertainty* when v and $V \equiv 0$ are given functions, i.e., $\bar{Q} \equiv 0$, $\tilde{V} \equiv \hat{V} \equiv 0$; (II) *under uncertainty* including the following three subcases: (II,i) *only additive uncertainty* ($V \equiv 0$); (II,ii) *only matrix uncertainty* ($\bar{\mathcal{Q}} \equiv 0$); (II,iii) *both ones.*

In [15–17], for cases (I) and (II,i) with controls (2), the following problem of *terminal target control synthesis under uncertainty* was investigated.

Problem 1. For the system (1), (2), (4), case (I) or (II,i), specify a *solvability set* $\mathcal{W}(\tau, \theta, \mathcal{M}) = \mathcal{W}(\tau)$ and a set-valued *feedback control strategy*[2] $u = \mathrm{u}(t, x)$, $\mathrm{u}(\cdot, \cdot) \in U^c_{\mathcal{R}}$, such that all solutions to the differential inclusion $\dot{x} \in A(t)x + \mathrm{u}(t, x) + \mathcal{Q}(t)$, $t \in T$, that start from any given position $\{\tau, x_\tau\}$, $x_\tau = x(\tau) \in \mathcal{W}(\tau, \theta, \mathcal{M})$, $\tau \in [0, \theta)$, would reach the terminal set $\mathcal{M}$ at time θ: $x(\theta) \in \mathcal{M}$.

[1] The normality condition $\|p^i\|_2 = 1$ may be omitted to simplify formulas.

[2] Here the class $U^c_{\mathcal{R}}$ of *feasible control strategies* is taken to consist of all convex compact-valued multifunctions $\mathrm{u}(t, x)$ that are measurable in t, upper semicontinuous in x, being restricted by $\mathrm{u}(t, x) \subseteq \mathcal{R}(t)$, $t \in T$. The condition $\mathrm{u}(\cdot, \cdot) \in U^c_{\mathcal{R}}$ ensures that the corresponding differential inclusion does have a solution.

The multivalued function $\mathcal{W}(t)$, $t \in T$, is known as a *solvability tube* $\mathcal{W}(\cdot)$.

The ellipsoidal synthesis was elaborated in [15,17] for solving Problem 1. In [8], the families of external $\mathcal{P}^+(\cdot)$ and internal $\mathcal{P}^-(\cdot)$ parallelotope-valued (shorter, *polyhedral*) estimates for $\mathcal{W}(\cdot)$ were introduced. The extremal aiming strategies of N.N. Krasovskii were used there. They were constructed in an analytical form on the base of a solution of some specific mathematical programming problem. Now let us consider two following problems, which concern all above cases of uncertainties. Unlike Problem 1, they involve single-valued control strategies. This is possible because our strategies will be continuous and even linear with respect to x. Moreover, they will be constructed in an explicit form.

Problem 2. For the system (1), (2), (4), (5), find a polyhedral tube $\mathcal{P}^-(t) = \mathcal{P}[p^-(t), \bar{P}^-(t)]$, $t \in T$, with $\mathcal{P}^-(\theta) = \mathcal{M}$, and find a corresponding feedback control strategy $u = u(t,x)$ such that $u(t,x) \in \mathcal{R}(t)$ for $x \in \mathcal{P}^-(t)$, $t \in T$, and each solution $x(\cdot)$ to the differential equation $\dot{x} = (A(t) + V(t))x + u(t,x) + v(t)$, $t \in T$, with $x(0) = x_0 \in \operatorname{int} \mathcal{P}^-(0)$ would be defined on T and would satisfy $x(t) \in \mathcal{P}^-(t)$, $t \in T$, whatever are $v(\cdot)$ and $V(\cdot)$ subjected to (4). Moreover, introduce a whole family of such tubes $\mathcal{P}^-(\cdot)$.

Problem 3. For the system (1), (3), (4), (5), find a polyhedral tube $\mathcal{P}^-(t) = \mathcal{P}[p^-(t), \bar{P}^-(t)]$, $t \in T$, with $\mathcal{P}^-(\theta) = \mathcal{M}$, and find a corresponding feedback control strategy $U = U(t,x)$ such that $U(t,x) \in \mathcal{U}(t)$ for $x \in \mathcal{P}^-(t)$, $t \in T$, and each solution $x(\cdot)$ to the differential equation

$$\dot{x} = (A(t) + U(t,x) + V(t))x + v(t), \qquad t \in T, \tag{6}$$

with $x(0) = x_0 \in \operatorname{int} \mathcal{P}^-(0)$ would be defined on T and would satisfy $x(t) \in \mathcal{P}^-(t)$, $t \in T$, whatever are $V(\cdot)$, $v(\cdot)$ subjected to (4). Introduce a family of such tubes.

3 Solutions to Problem 2

First, let us consider the following ODE system for $\mathcal{P}^-(t) = \mathcal{P}[p^-(t), \bar{P}^-(t)]$:

$$\frac{dp^-}{dt} = (A(t) + \tilde{V}(t))p^- + r(t) + q(t), \quad p^-(\theta) = p_{\mathrm{f}}; \tag{7}$$

$$\begin{gathered}\frac{d\bar{P}^-}{dt} = (A(t) + \tilde{V}(t))\bar{P}^- + \bar{P}^- \operatorname{diag} \beta(t, \bar{P}^-) + \bar{R}(t)\, \Gamma(t) + \bar{P}^- \operatorname{diag} \gamma(t, \bar{P}^-), \\ \beta(t, \bar{P}^-) = \max\{\operatorname{Abs}((\bar{P}^-)^{-1})\, \hat{V}(t) \operatorname{Abs}(p^-(t) + \bar{P}^- \xi) \mid \xi \in \mathbb{E}(\mathcal{C})\}, \\ \gamma(t, \bar{P}^-) = \operatorname{Abs}((\bar{P}^-)^{-1} \bar{Q}(t))\, \mathrm{e}, \qquad \bar{P}^-(\theta) = \bar{P}_{\mathrm{f}}.\end{gathered} \tag{8}$$

Here (and below) the operation of maximum is understood componentwise, $\mathbb{E}(\mathcal{C})$ denotes the set of all vertices of $\mathcal{C} = \mathcal{P}(0, I, \mathrm{e})$ (i.e., points $\xi \in \mathbb{R}^n$ with $\xi_j \in \{-1, 1\}$); $\Gamma(t) \in \mathbb{R}^{n_1 \times n}$ is an arbitrary Lebesgue measurable matrix function satisfying $\Gamma(t) \in \mathcal{G}$, a.e. $t \in T$, where $\mathcal{G} = \{\Gamma = \{\gamma_i^j\} \in \mathbb{R}^{n_1 \times n} \mid \|\Gamma\| \leq 1\}$,

$\|\Gamma\| = \max_{1\le i\le n_1} \sum_{j=1}^{n} |\gamma_i^j|$. Let $\mathbb{G}$ be the set of all such functions $\Gamma(\cdot)$. Let us consider the following control strategy, being connected with $\mathcal{P}^-(\cdot)$ from (7), (8):

$$u(t,x) = r(t) + \bar{R}(t)\Gamma(t)\bar{P}^-(t)^{-1}(x - p^-(t)). \tag{9}$$

Theorem 1. *We consider the system (1), (2), (4), (5), where* $\det \bar{P}_\mathrm{f} \neq 0$. *Let* $\Gamma(\cdot) \in \mathbb{G}$. *Then the system (7), (8) has a unique solution* $(p^-(\cdot), \bar{P}^-(\cdot))$ *at least on some subinterval* $T_1 = [\tau_1, \theta] \subseteq T$, *where* $0 \le \tau_1 < \theta$. *If* $T_1 = T$ *and we have* $\det \bar{P}^-(t) \neq 0$, $t \in T$, *then the tube* $\mathcal{P}^-(\cdot)$ *and the control strategy (9) give a particular solution to Problem 2; in cases (I), (II,i), all solutions* $x(\cdot)$ *with* $x(0) \in \mathcal{P}^-(0)$ *(not only with* $x(0) \in \operatorname{int} \mathcal{P}^-(0)$*) generated by (9) satisfy* $x(t) \in \mathcal{P}^-(t)$, $t \in T$.

The scheme of the proof is similar to the proof of Theorem 2 (see below).

Theorem 1 describes the whole family of tubes $\mathcal{P}^-(\cdot)$, where $\Gamma(\cdot)$ serves as a parameter. Thus the set $\mathcal{W}^0 = \bigcup\{\operatorname{int} \mathcal{P}^-(0) |\, \Gamma(\cdot) \in \mathbb{G}$ such that $\det \mathcal{P}^-(t) \neq 0, t \in T\}$ (or, in cases (I), (II,i), the analogous set $\mathcal{W}^0 = \bigcup \mathcal{P}^-(0)$) provides the set of initial positions which can be steered to the terminal set $\mathcal{M}$ during the time θ by solving Problem 2. But, generally speaking, it is not true that $\det \mathcal{P}^-(0) \neq 0$ or even $\mathcal{P}^-(0) \neq \emptyset$ for each $\Gamma(\cdot) \in \mathbb{G}$. For cases (I), (II,i), the above family of the tubes $\mathcal{P}^-(\cdot)$ coincides with the family of internal estimates for $\mathcal{W}(\cdot)$ introduced in [8]. It follows from [8,11] that for the case (I) we have $T_1 = T$ for each $\Gamma(\cdot) \in \mathbb{G}$ and $\mathcal{W}(0) = \bigcup\{\mathcal{P}^-(0) |\, \Gamma(\cdot) \in \mathbb{G}\}$. But we can not conclude from here that $\mathcal{W}^0 = \mathcal{W}(0)$. The attractive property of the control strategies (9) is their explicit form.

Remark 1. One of the heuristic ways to construct the parameter $\Gamma(\cdot)$ is to apply arguments of a "local" volume optimization similarly to [8] (see also Remark 2 below). Namely, assuming $\det \bar{P}_\mathrm{f} > 0$ and introducing a grid T_N of times $\tau_k = kh_N$, $k = 0, \ldots, N$, $h_N = \theta N^{-1}$, we can construct the piecewise constant function $\Gamma(t) \equiv \Gamma(\tau_k) \in \operatorname{Argmin}_{\Gamma\in\mathcal{G}} \operatorname{tr}(\bar{P}^-(\tau_k)^{-1}\bar{R}(\tau_k)\Gamma)$, $t \in (\tau_{k-1}, \tau_k]$, $k = N, \ldots, 1$.

For case (I), we can use, similarly to [11], minimization over Γ that satisfy $\Gamma \in \mathcal{G}$ and some constraints introduced to produce tight estimates $\mathcal{P}^-(t)$ for $\mathcal{W}(t)$. Solutions of both optimization problems are known in the explicit form [8,11].

4 Solutions to Problem 3

Let us consider the following ODE system for $\mathcal{P}^-(t) = \mathcal{P}[p^-(t), \bar{P}^-(t)]$:

$$\frac{dp^-}{dt} = (A(t) + \tilde{U}(t) + \tilde{V}(t))p^- + q(t), \quad p^-(\theta) = p_\mathrm{f}; \tag{10}$$

$$\begin{aligned}
&\frac{d\bar{P}^-}{dt} = (A(t) + \tilde{U}(t) + \tilde{V}(t))\bar{P}^- - \operatorname{diag}\alpha(t,\bar{P}^-)\bar{P}^- + \bar{P}^-\operatorname{diag}(\beta(t,\bar{P}^-) + \gamma(t,\bar{P}^-)),\\
&\quad \alpha_i(t,\bar{P}^-) = \alpha_i(t,\bar{P}^-;J(t)) = \hat{u}_i^{j_i}(t)\,\eta_{j_i}(t,\bar{P}^-)(\mathrm{e}^{i\top}(\operatorname{Abs}\bar{P}^-)\,\mathrm{e})^{-1}, \quad i=1,\dots,n,\\
&\quad \eta(t,\bar{P}^-) = \max\{0, \operatorname{Abs} p^-(t) - (\operatorname{Abs}\bar{P}^-)\mathrm{e}\},\\
&\quad \beta(t,\bar{P}^-) = \max\{\operatorname{Abs}((\bar{P}^-)^{-1})\,\hat{V}(t)\operatorname{Abs}(p^-(t) + \bar{P}^-\xi) \mid \xi\in\mathbb{E}(\mathcal{C})\},\\
&\quad \gamma(t,\bar{P}^-) = \operatorname{Abs}((\bar{P}^-)^{-1}\bar{Q}(t))\,\mathrm{e}, \qquad \bar{P}^-(\theta) = \bar{P}_{\mathrm{f}},
\end{aligned} \tag{11}$$

where $\hat{u}_i^j$ stand for elements of $\hat{U}$. Here $J = \{j_1,\dots,j_n\}$ is an arbitrary permutation of numbers $\{1,\dots,n\}$ or even a measurable vector function with values $J(t)$ being arbitrary permutations. Let $\mathbb{J}$ be the set of all such functions $J(\cdot)$. Let us consider the control strategy connected with $\mathcal{P}^-(\cdot)$ from (10), (11):

$$\mathrm{e}^{i\top}U(t,x) = \begin{cases} \mathrm{e}^{i\top}\tilde{U}(t) - \alpha_i(t,\bar{P}^-(t))(x_i - p_i^-(t))(x_{j_i})^{-1}\mathrm{e}^{j_i\top} & \text{if } x_{j_i}\neq 0,\\ \mathrm{e}^{i\top}\tilde{U}(t) \quad \text{if } x_{j_i} = 0, & i=1,\dots,n. \end{cases} \tag{12}$$

Theorem 2. *We consider the system (1), (3)–(5), where* $\det\bar{P}_{\mathrm{f}}\neq 0$. *Let* $J(\cdot)\in\mathbb{J}$. *Then the system (10), (11) has a unique solution* $(p^-(\cdot),\bar{P}^-(\cdot))$ *at least on some subinterval* $T_1 = [\tau_1,\theta]\subseteq T$, *where* $0\le\tau_1<\theta$. *If* $T_1 = T$ *and we have* $\det\bar{P}^-(t)\neq 0$, $t\in T$, *then the tube* $\mathcal{P}^-(\cdot)$ *and the control strategy (12) give a particular solution to Problem 3; in cases (I), (II,i), all solutions* $x(\cdot)$ *to (6) with* $x(0)\in\mathcal{P}^-(0)$ *(not only with* $x(0)\in\operatorname{int}\mathcal{P}^-(0)$*) satisfy* $x(t)\in\mathcal{P}^-(t)$, $t\in T$.

Proof. Here we give a sketch. First, it can be checked that the strategy (12) acts for $x\in\mathcal{P}^-(t)$ according to the rule $U(t,x)\,x = \tilde{U}(t)\,x - \operatorname{diag}\alpha(t,\bar{P}^-(t))\cdot(x - p^-(t))$. Existence and uniqueness of the solution follow from the known results similarly to [8,10]. Let $x_0\in\operatorname{int}\mathcal{P}^-(0)$ ($x_0\in\mathcal{P}^-(0)$ for cases (I) and (II,i)). Let $x(\cdot)$ be the solution of (6) that corresponds to $x(0) = x_0$ (i.e., $x(0) = p^-(0) + \bar{P}^-(0)\zeta_0$, where $\|\zeta_0\|_\infty < 1$ (respectively, $\|\zeta_0\|_\infty\le 1$)), to the control $U(t,x)$ from (12), and to arbitrary admissible functions $v(\cdot)$ (such that $v(t) = q(t) + \bar{Q}(t)\chi(t)$, $\|\chi(t)\|_\infty\le 1$) and $V(\cdot)$ (which satisfies (4)). Let us represent $x(t) - p^-(t)$ in the form $x(t) - p^-(t) = \bar{P}^-(t)\zeta(t)$. Then we have $\frac{d}{dt}\zeta = -(\bar{P}^-)^{-1}(\frac{d}{dt}\bar{P}^-)\zeta + (\bar{P}^-)^{-1}\frac{d}{dt}(x-p^-)$ for the above function ζ. Taking into account (11) and the relation $\frac{d}{dt}(x-p^-) = (A+\tilde{U}+\tilde{V})(x-p^-) - (\operatorname{diag}\alpha)(x-p^-) + (V-\tilde{V})x + v - q$, which follows from (6), (10), (12), it is not difficult to see that $\dot{\zeta} = -(\operatorname{diag}\beta + \operatorname{diag}\gamma)\zeta + (\bar{P}^-)^{-1}((V-\tilde{V})x + v - q)$. Let us denote $b(t) = \beta(t,\bar{P}^-(t)) + \gamma(t,\bar{P}^-(t))$, $c(t,\zeta) = \bar{P}^-(t)^{-1}((V(t)-\tilde{V}(t))\cdot(p^-(t) + \bar{P}^-(t)\zeta) + \bar{Q}(t)\chi(t))$. Then, using (4), (5), we have

$$\begin{aligned}
&\dot{\zeta}_i = -b_i(t)\zeta_i + c_i(t,\zeta), \quad i=1,\dots,n, \qquad \zeta(0) = \zeta_0;\\
&b(t)\ge 0, \quad \operatorname{Abs} c(t,\zeta)\le b(t) \quad \text{for} \quad \zeta\in\mathcal{C} = \mathcal{P}(0,I,\mathrm{e}).
\end{aligned} \tag{13}$$

It is not difficult to check that if $\zeta(\cdot)$ satisfies (13) and $\zeta_0\in\operatorname{int}\mathcal{C}$, then $\zeta(t)\in\operatorname{int}\mathcal{C}$, $t\in T$; if $\zeta_0\in\mathcal{C}$ and, in addition, $c(t,\zeta)\equiv c(t)$ (i.e., does not depend on ζ), then

$\zeta(t) \in \mathcal{C}$, $t \in T$. Thus we obtain $x(t) \in \mathcal{P}^-(t)$, $t \in T$. Also we have $\mathrm{Abs}\,(U(t,x) - \tilde{U}(t)) \le \hat{U}(t)$ for $x \in \mathcal{P}^-(t)$ because for such x we have $|\alpha_i(x_i - p_i^-)(x_{j_i})^{-1}| \le \hat{u}_i^{j_i}$, $i = 1, \ldots, n$ (this can be obtained by simple estimates). □

Theorem 2 describes the family of tubes $\mathcal{P}^-(\cdot)$, where $J(\cdot)$ serves as a parameter. Thus the set $\mathcal{W}^0 = \bigcup\{\mathrm{int}\,\mathcal{P}^-(0) |\, J(\cdot) \in \mathbb{J}$ such that $\det \mathcal{P}^-(t) \ne 0,\ t \in T\}$ provides the set of x_0 that can be steered to $\mathcal{M}$ during the time θ by solving Problem 3. However it is not true that $\det \mathcal{P}^-(0) \ne 0$ or $\mathcal{P}^-(0) \ne \emptyset$ for each $J(\cdot) \in \mathbb{J}$.

Remark 2. One of the ways of constructing $J(\cdot)$ is to apply arguments of a "local" volume optimization similarly to [9,10]. Namely, assume, without loss of generality, that $\det \bar{P}_{\mathrm{f}} > 0$. Fix a natural number N and introduce a grid T_N of times $\tau_k = kh_N$, $k = 0, \ldots, N$, $h_N = \theta N^{-1}$. Integrating the system (10), (11) from right to left, let us, for each $\tau \in T_N$, solve the optimization problem which is to maximize $\sum_{i=1}^n \alpha_i(\tau, \bar{P}^-(\tau); J)$ over all possible permutations $J = \{j_1, \ldots, j_n\}$. This is equivalent to finding the maximal possible velocity of increasing (from right to left) $\det \bar{P}^-(\tau)$ (therefore $\mathrm{vol}\,\mathcal{P}^-(\tau)$) at time τ, by the choice of the value J, when the value $\bar{P}^-(\tau)$ has already been found. Thus we can sequentially construct the piecewise constant function $J(t) \equiv J(\tau_k) \in \mathrm{Argmax}_J \sum_{i=1}^n \alpha_i(\tau_k, \bar{P}^-(\tau_k); J)$, $t \in (\tau_{k-1}, \tau_k]$, $k = N, \ldots, 1$, and find $\bar{P}^-(\cdot)$.

5 Control Synthesis for Discrete-Time Systems

Now let us briefly consider a problem of control synthesis, similar to Problem 3, for discrete-time systems. This is of independent interest and also may be useful for constructing difference schemes for solving the system (10), (11). The analog of Problem 2 can be considered in a similar way.

Consider the controlled discrete-time system with a given terminal set $\mathcal{M}$:

$$\begin{gathered} x[k] = (A[k] + U[k] + V[k])\, x[k-1] + v[k], \quad k = 1, \ldots, N, \\ x[N] \in \mathcal{M} = \mathcal{P}[p_{\mathrm{f}}, \bar{P}_{\mathrm{f}}], \quad \det \bar{P}_{\mathrm{f}} \ne 0, \end{gathered} \tag{14}$$

$$U[k] \in \mathcal{U}[k] = \{U |\, \mathrm{Abs}\,(U - \tilde{U}[k]) \le \hat{U}[k]\},\ V[k] \in \{V |\, \mathrm{Abs}\,(V - \tilde{V}[k]) \le \hat{V}[k]\}, \tag{15}$$

$$v[k] \in \mathcal{Q}[k] = \mathcal{P}[q[k], \bar{Q}[k]], \quad k = 1, \ldots, N. \tag{16}$$

Problem 4. Find a polyhedral tube $\mathcal{P}^-[k] = \mathcal{P}[p^-[k], \bar{P}^-[k]]$, $k = 1, \ldots, N$, with $\mathcal{P}^-[N] = \mathcal{M}$, and find a corresponding feedback control strategy $U = U[k, x]$ such that $U[k, x] \in \mathcal{U}[k]$ for $x \in \mathcal{P}^-[k-1]$, $k = 1, \ldots, N$, and each solution $x[\cdot]$ to the equation $x[k] = (A[k] + U[k, x[k-1]] + V[k]) \cdot x[k-1] + v[k]$, $k = 1, \ldots, N$, with $x[0] = x_0 \in \mathcal{P}^-[0]$ would satisfy $x[k] \in \mathcal{P}^-[k]$, $k = 1, \ldots, N$, whatever are $V[\cdot]$ and $v[\cdot]$ subjected to (15), (16). Introduce a family of such tubes $\mathcal{P}^-[\cdot]$.

Let us consider the following system of relations for $\mathcal{P}^-[k] = \mathcal{P}[p^-[k], \bar{P}^-[k]]$:

$$p^-[k-1] = B[k]^{-1}(p^-[k] - q[k]),\ B[k] = A[k] + \tilde{U}[k] + \tilde{V}[k],\ k = N, \ldots, 1,\ p^-[N] = p_{\mathrm{f}}, \tag{17}$$

$$\bar{P}^-[k-1] = H[k, \bar{P}^-[k-1]], \quad k = N, \dots, 1, \qquad \bar{P}^-[N] = \bar{P}_{\mathrm{f}}, \tag{18}$$

$$\begin{gathered} H[k,P] = (B[k] - \operatorname{diag}\alpha[k,P])^{-1}\bar{P}^-[k]\operatorname{diag}(\mathrm{e} - \beta[k,P] - \gamma[k]), \\ \alpha_i[k,P] = \alpha_i[k,P;J[k]] = \hat{u}_i^{j_i}[k]\,\eta_{j_i}[k,P]\,(\mathrm{e}^{i\top}(\operatorname{Abs}P)\,\mathrm{e})^{-1}, \quad i = 1,\dots,n, \\ \eta[k,P] = \max\{0, \operatorname{Abs}p^-[k-1] - (\operatorname{Abs}P)\mathrm{e}\}, \\ \beta[k,P] = \max\{\operatorname{Abs}(\bar{P}^-[k]^{-1})\,\hat{V}[k]\operatorname{Abs}(p^-[k-1] + P\xi) \mid \xi \in \mathbb{E}(\mathcal{C})\}, \\ \gamma[k] = (\operatorname{Abs}(\bar{P}^-[k]^{-1}\bar{Q}[k]))\mathrm{e}, \quad k = N,\dots,1. \end{gathered} \tag{19}$$

Note that (17) is the system of explicit recurrent relations while (18)–(19) is the system of implicit ones, i.e., for any time step $k \in \{N,\dots,1\}$, we need to solve the system of nonlinear equations with respect to the unknown matrix $P = P^-[k-1]$.

Theorem 3. *In the system (14)–(16), let* $\det\bar{P}_{\mathrm{f}} \neq 0$ *and all* $\det B[k] \neq 0$. *Let* $J[k] = \{j_1[k],\dots,j_n[k]\}$ *be arbitrary permutations of numbers* $\{1,\dots,n\}$, $k = N,\dots,1$, *and the system (17)–(19) has a solution* $(p^-[\cdot], \bar{P}^-[\cdot])$ *such that we obtain* $\det\bar{P}^-[k] \neq 0$ *and* $\mathrm{e} - \beta[k, \bar{P}^-[k-1]] - \gamma[k] > 0$, $k = N,\dots,1$. *Then the tube* $\mathcal{P}^-[\cdot]$ *and the control strategy which acts according to the following rule*

$$U[k,x]\,x = \tilde{U}[k]\,x - \operatorname{diag}\alpha[k,\bar{P}^-[k-1];J[k]](x - p^-[k-1]), \quad k = 1,\dots,N, \tag{20}$$

(a formula similar to (12) is true), gives a particular solution to Problem 4.

Proof. We give a sketch following the scheme of the proof of Theorem 2 and keeping the similar notation. Let $x[\cdot]$ corresponds to $x[0] = x_0 \in \mathcal{P}^-[0]$, i.e., $x[0] = p^-[0] + \bar{P}^-[0]\zeta_0$, where $\|\zeta_0\|_\infty \leq 1$. Let us represent $x[k]$ in the form $x[k] = p^-[k] + \bar{P}^-[k]\zeta[k]$, $k = 0,\dots,N$. The proof is by induction on the time step k. Let we already have $x[k-1] \in \mathcal{P}^-[k-1]$. Then it follows from (14), (17) that

$$x[k] = p^-[k] + B[k]\bar{P}^-[k-1]\zeta[k-1] + (U[k,x[k-1]] - \tilde{U}[k] + \Delta V[k])x[k-1] + v[k] - q[k],$$

where $\Delta V[k] = V[k] - \tilde{V}[k]$. Taking into account (20), we obtain

$$\zeta[k] = \bar{P}^-[k]^{-1}(B[k] - \operatorname{diag}\alpha[k,\bar{P}^-[k-1]])\bar{P}^-[k-1]\zeta[k-1] + c[k,x[k-1]],$$

$c[k,x] = \bar{P}^-[k]^{-1}\Delta V[k]\,x + \bar{P}^-[k]^{-1}\bar{Q}[k]\chi[k]$. Using (18), (19), (15), (16), we have

$$\begin{gathered} \zeta[k] = \operatorname{diag}(\mathrm{e} - \beta[k,\bar{P}^-[k-1]] - \gamma[k])\zeta[k-1] + c[k,x[k-1]]; \\ \operatorname{Abs}c[k,x] \leq \beta[k,\bar{P}^-[k-1]] + \gamma[k] \quad \text{for} \quad x \in \mathcal{P}^-[k-1]. \end{gathered}$$

It is not difficult to see that if $\|\zeta[k-1]\|_\infty \leq 1$ and $\mathrm{e} - \beta[k,\bar{P}^-[k-1]] - \gamma[k] \geq 0$, then $\|\zeta[k]\|_\infty \leq 1$. Thus we obtain the desired inclusion $x[k] \in \mathcal{P}^-[k]$. Also it is not difficult to see that $\operatorname{Abs}(U[k,x] - \tilde{U}[k]) \leq \hat{U}[k]$ for $x \in \mathcal{P}^-[k-1]$. □

Remark 3. Let the system (14)–(16) be obtained by the Euler approximations of (1), (3)–(5) with the same $\mathcal{M}$, $A[k] = I + h_N A(t_{k-1})$, $\tilde{U}[k] = h_N\tilde{U}(t_{k-1})$,

$\hat{U}[k] = h_N \hat{U}(t_{k-1})$, $\tilde{V}[k] = h_N \tilde{V}(t_{k-1})$, $\hat{V}[k] = h_N \hat{V}(t_{k-1})$, $\mathcal{Q}[k] = h_N \mathcal{Q}(t_{k-1})$, $t_k = kh_N$, $h_N = \theta N^{-1}$. Let, for a fixed k, $\det \bar{P}^-[k] \neq 0$ and the time step h_N be sufficient small. Then the operator $H[k, P]$ is contractive in some domain $\mathcal{D}[k] = \{P | \, \|P - \bar{P}^-[k]\| \leq \delta[k]\}$, i.e., $\|H[k, P^1] - H[k, P^2]\| \leq L\|P^1 - P^2\|$ for any $P^1, P^2 \in \mathcal{D}[k]$, where $L = L[k] \in (0, 1)$, and therefore [3, p. 319] the equation $P = H[k, P]$ from (18), (19) has a solution $P = \bar{P}^-[k-1]$, which can be found by the simple iteration $P^{l+1} = H[k, P^l]$, $l = 0, 1, \ldots$, starting from $P^0 = \bar{P}^-[k]$, and we have $\|P^l - P\| \leq L^l(1-L)^{-1}\|P^1 - P^0\|$. Also, the relation $\gamma[k] + \beta[k, \bar{P}^-[k-1]] < \mathrm{e}$ is satisfied. But certainly we can not derive from here the existence of nonsingular matrices $\bar{P}^-[k]$ for all $k = N, \ldots, 1$ because the value of such "small" h_N depends on k.

6 Examples

We consider model examples for Problem 3. For computations we use the Euler approximations (see Remark 3) with $N = 200$. Let $\mathcal{M} = \mathcal{P}((1,1)^\top, I, (0.1, 0.1)^\top)$,

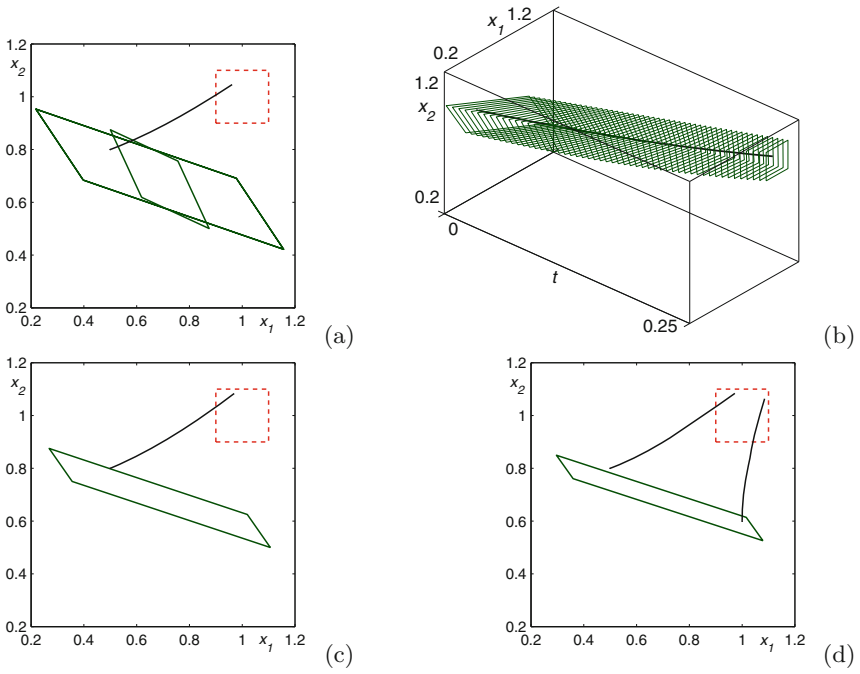


Fig. 1. Examples of polyhedral control synthesis for Problem 3 ($n = 2$). (a) Case (I): the set $\mathcal{M}$ (dash line), two parallelotopes $\mathcal{P}^-[0]$ and the controlled trajectory for $x_0 = (0.5, 0.8)^\top$. (b) Case (I): the tube $\mathcal{P}^-[\cdot]$ and the controlled trajectory. (c) Case (II,ii): $\mathcal{M}$, $\mathcal{P}^-[0]$ corresponding to $J[\cdot]$ from Remark 2, and the controlled trajectory for $x_0 = (0.5, 0.8)^\top$. (d) Case (II,iii): $\mathcal{M}$, $\mathcal{P}^-[0]$ corresponding to $J[\cdot]$ from Remark 2, and controlled trajectories for two initial points $x_0 = (0.5, 0.8)^\top$ and $x_0 = (1, 0.6)^\top$.

$A(t) \equiv \begin{bmatrix} -0.5 & 0 \\ 0 & -0.5 \end{bmatrix}$, $\tilde{U}(t) \equiv \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix}$, $\hat{U}(t) \equiv \begin{bmatrix} 0 & 1.5 \\ 0 & 0 \end{bmatrix}$, $\tilde{V}(t) \equiv \begin{bmatrix} 0 & 0 \\ 2 & 0 \end{bmatrix}$, $\hat{V}(t) \equiv \begin{bmatrix} 0 & 0 \\ 0.2 & 0 \end{bmatrix}$ or $\hat{V}(t) \equiv 0$, $\mathcal{Q}(t) \equiv \mathcal{P}(0, I, 0)$ or $\mathcal{Q}(t) \equiv \mathcal{P}(0, I, (0.05, 0.05)^\top)$, $\theta = 0.25$. We consider 3 cases: (I), (II,ii), and (II,iii). The results are presented in Fig. 1. In the second example, we put the realization $V(t) \equiv \tilde{V}(t) + \hat{V}(t)$; in the third one we presume $V(\cdot)$ to be the same and $v(\cdot)$ to be some extremal bang-bang type disturbance [17, p. 234], where the length of intervals of constancy of $v(t)$ is equal to $\theta/4$.

All the presented trajectories reach the target set $\mathcal{M}$ including the one with a small violation of the inclusion $x_0 \in \mathcal{P}^-[0]$, though if $x_0 \notin \mathcal{P}^-[0]$, then there is not guarantee that the trajectory can be steered into the target set $\mathcal{M}$ by using the control strategy (12) under any disturbances.

References

1. Anan'evskii, I.M., Anokhin, N.V., Ovseevich, A.I.: Synthesis of a bounded control for linear dynamical systems using the general Lyapunov function. Dokl. Math. **82**(2), 831–834 (2010)
2. Baier, R., Lempio, F.: Computing Aumanns integral. In: Kurzhanski, A.B., Veliov, V.M. (eds.) Modeling Techniques for Uncertain Systems (Sopron, 1992). Progress in Systems and Control Theory, vol. 18, pp. 71–92. Birkhäuser, Boston (1994)
3. Bakhvalov, N.S., Zhidkov, N.P., Kobel'kov, G.M.: Numerical Methods. Nauka, Moscow (1987). (Russian)
4. Chernousko, F.L.: State Estimation for Dynamic Systems. CRS Press, Boca Raton (1994)
5. Filippova, T.F.: Trajectory tubes of nonlinear differential inclusions and state estimation problems. J. Concr. Appl. Math. **8**(3), 454–469 (2010)
6. Filippova, T.F., Lisin, D.V.: On the estimation of trajectory tubes of differential inclusions. Proc. Steklov Inst. Math. Suppl. **2**, S28–S37 (2000)
7. Gusev, M.I.: External estimates of the reachability sets of nonlinear controlled systems. Autom. Remote Control **73**(3), 450–461 (2012)
8. Kostousova, E.K.: Control synthesis via parallelotopes: optimization and parallel computations. Optim. Methods Softw. **14**(4), 267–310 (2001)
9. Kostousova, E.K.: State estimation for control systems with a multiplicative uncertainty through polyhedral techniques. In: Hömberg, D., Tröltzsch, F. (eds.) System Modeling and Optimization. IFIP AICT, vol. 391, pp. 165–176. Springer, Heidelberg (2013)
10. Kostousova, E.K.: On polyhedral estimates for reachable sets of differential systems with bilinear uncertainty. Trudy Instituta Matematiki i Mekhaniki UrO RAN **18**(4), 195–210 (2012). (Russian)
11. Kostousova, E.K.: On tight polyhedral estimates for reachable sets of linear differential systems. In: 9th International Conference on Mathematical Problems in Engineering, Aerospace and Sciences: ICNPAA 2012, Vienna, Austria, July 10–14, 2012. AIP Conf. Proc. **1493** (2012). doi:http://dx.doi.org/10.1063/1.4765545
12. 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)

13. Krasovskii, N.N., Subbotin, A.I.: Positional Differential Games. Nauka, Moscow (1974). (Russian)
14. Kuntsevich, V.M., Kurzhanski, A.B.: Calculation and control of attainability sets for linear and certain classes of nonlinear discrete systems. J. Autom. Inf. Sci. **42**(1), 1–18 (2010)
15. Kurzhanskii, A.B., Mel'nikov, N.B.: On the problem of the synthesis: the Pontryagin alternating integral and the Hamilton-Jacobi equation. Sb. Math. **191**(6), 849–882 (2000)
16. Kurzhanski, A.B., Nikonov, O.I.: On the problem of synthesizing control strategies: evolution equations and set-valued integration. Soviet Math. Dokl. **41**(2), 300–305 (1990)
17. Kurzhanski, A.B., Vályi, I.: Ellipsoidal Calculus for Estimation and Control. Birkhäuser, Boston (1997)
18. Kurzhanski, A.B., Varaiya, P.: On ellipsoidal techniques for reachability analysis. Part I: External approximations. Part II: Internal approximations. Box-valued constraints. Optim. Methods Softw. **17**(2), 177–237 (2002)
19. Mazurenko, S.S.: A differential equation for the gauge function of the star-shaped attainability set of a differential inclusion. Dokl. Math. **86**(1), 476–479 (2012)
20. Polyak, B.T., Scherbakov, P.S.: Robust Stability and Control. Nauka, Moscow (2002). (Russian)
21. 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. Syst. Sci. Int. **33**(6), 127–139 (1995)
22. Veliov, V.M.: Second order discrete approximations to strongly convex differential inclusions. Syst. Control Lett. **13**(3), 263–269 (1989)

Application of the Fenchel Theorem to the Obstacle Problem

Diana R. Merluşcă(✉)

Institute of Mathematics of the Romanian Academy, Bucharest, Romania
dianam1985@yahoo.com

Abstract. In this paper we apply a duality algorithm to the general obstacle problem for second order operators. We reduce the problem to the null obstacle case and we solve it by using an algorithm based on a dual approximate problem. This method generates a quadratic minimization problem, which is easy to implement numerically. The convergence properties and the numerical results show that the algorithm is working properly for any admissible obstacle.

Keywords: Obstacle problem · Duality · Approximate problem

1 Introduction

The obstacle problem is a very well studied subject. Many methods have been applied for solving it. For instance, Glowinski [6] used the finite element method for solving the null obstacle problem, while Barbu and Precupanu [2] studied the problem from the duality point of view. The general obstacle problem is also intensively studied for its wide range of applications in mechanics and physics. We refer here to Rodrigues [14], Caffarelli and Friedman [4], Duvaut and Lions [5] and Ciarlet [3].

We extend here the duality method developed in the articles Merluşcă [8,9], by the application of the Fenchel theorem to the obstacle problem. We discuss the general obstacle problem in Sect. 2. We reduce the problem to the null obstacle case and we compute the solutions using the duality method (Merluşcă [9]). In Merluşcă [10], the case of the fourth order obstacle problem was analysed. In Sect. 3, we apply this technique in numerical examples for one dimensional problems and the obtained results are very accurate. Finally, we mention the works of Neittaanmaki, Sprekels and Tiba [12] and Sprekels and Tiba [15], where a related duality approach was used in the study of Kirchhoff-Love arches and explicit solutions were obtained.

This paper is supported by the Sectorial Operational Programme Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under contract number SOP HRD/107/1.5/S/82514.

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 181–188, 2014.
DOI: 10.1007/978-3-662-45504-3_17

2 The General Obstacle Problem for Second Order Equations

We consider the following obstacle problem

$$\min\left\{\frac{1}{2}\int_\Omega |\nabla y|^2 - \int_\Omega fy \ : \ y \in K_\psi\right\}, \tag{1}$$

where $K_\psi = \{y \in H_0^1(\Omega) : y \geq \psi\}$, $\psi \in H^1(\Omega)$ is such that $\psi|_{\partial\Omega} \leq 0$ and $f \in L^2(\Omega)$.

It is known that the unique solution of problem (1) is an element in $H^2(\Omega)$ (Theorem 2.5, [1]).

Lemma 1. *Let y_ψ be the solution of the problem (1) and $\hat{y}$ the solution of the problem*

$$\begin{aligned} -\Delta\hat{y} &= f, \ \text{ on } \ \Omega, \\ \hat{y} &= 0, \quad \text{on } \ \partial\Omega, \end{aligned} \tag{2}$$

then $y_\psi \geq \hat{y}$ almost everywhere on Ω.

The problem (1) in which we replace ψ by $\hat{\psi} = \max\{\hat{y}, \psi\} \in H_0^1(\Omega)$ has the same solution y_ψ.

Proof. Denoting $\beta \subset \mathbb{R} \times \mathbb{R}$ a maximal monotone operator defined by

$$\beta(z) = \begin{cases}]-\infty, 0], & z = 0, \\ 0, & z > 0\,, \\ \emptyset, & z < 0. \end{cases}$$

we rewrite (1) as

$$-\Delta y_\psi + \beta(y_\psi - \psi) \ni f \quad \text{in } \ \Omega. \tag{3}$$

Then, since $y_\psi \in H^2(\Omega)$, $\beta(y_\psi - \psi) \in L^2(\Omega)$ and $\beta(y_\psi - \psi) \leq 0$ a.e. on Ω. By a comparison of (2) and (3), we obtain that $y_\psi \geq \hat{y}$ a. e. on Ω.

We denote $\hat{K} = \{y \in H_0^1(\Omega) : y \geq \hat{\psi}\}$. Then $y_\psi \in \hat{K}$, $\Delta y_\psi + f \leq 0$ a.e. on Ω.

For every $v \in \hat{K}$, we compute

$$\begin{aligned} \int_\Omega (\Delta y_\psi + f)(v - y_\psi) &= \int_\Omega (\Delta y_\psi + f)(\hat{\psi} - y_\psi) + \int_\Omega (\Delta y_\psi + f)(v - \hat{\psi}) \\ &\leq \int_\Omega (\Delta y_\psi + f)(\hat{\psi} - y_\psi) = 0. \end{aligned}$$

The last equality is due to the classical formulation of the obstacle problem (the complementarity property)

$$\begin{aligned} -\Delta y_\psi &= f, \quad \text{in } \ \Omega^+ = \{y_\psi \in \Omega : y_\psi(x) > \psi(x)\}, \\ -\Delta y_\psi &\geq f, \quad \text{in } \ \Omega \setminus \Omega^+ = \{y_\psi \in \Omega : y_\psi(x) = \psi(x)\}, \\ y_\psi &= \psi \ \text{ and } \ \frac{\partial y_\psi}{\partial n} = \frac{\partial \psi}{\partial n}, \quad \text{on } \ \partial\Omega^+ \cap \Omega, \\ y_\psi &= 0 \quad \text{on } \ \partial\Omega. \end{aligned}$$

and to the fact that $y_\psi(x) = \psi(x)$ means that $\hat{y}(x) \leq \psi(x)$ and that yields that $y_\psi(x) = \hat{\psi}(x)$.

Then integrating by parts we get

$$\int_\Omega \nabla y_\psi \nabla(v - y_\psi) \leq \int_\Omega f(v - y_\psi), \quad \forall v \geq \hat{\psi}.$$

Remark 1. Lemma 1 is known (see Murea and Tiba [11]) and we indicate its proof for easy reference.

The null obstacle problem that we use is

$$\min_{y \in K_0} \left\{ \frac{1}{2} \int_\Omega |\nabla y|^2 - \int_\Omega fy + \int_\Omega \nabla \hat{\psi} \nabla y \right\}. \tag{4}$$

Let y_0 be the unique solution of problem (4).

Proposition 1. *Then the solution of the problem* (1) *can be computed by just adding* $\hat{\psi}$*, i.e.*

$$y_\psi = y_0 + \hat{\psi}. \tag{5}$$

Proof. The weak formulation of problem (4) is given by the form

$$\int_\Omega \nabla y_0 \nabla(y_0 - v) \leq \int_\Omega f(y_0 - v) - \int_\Omega \nabla \hat{\psi} \nabla(y_0 - v), \quad \forall v \in K_0.$$

We translate the problem by adding $\hat{\psi}$. Then, for every $v \in K_0$, we get $v + \hat{\psi} \geq \hat{\psi} \geq \psi$. With this translations from the variational inequality we obtain

$$\int_\Omega \nabla(y_0 + \hat{\psi})(\nabla(y_0 + \hat{\psi}) - \nabla(\hat{\psi} + v)) \leq \int_\Omega f(y_0 + \hat{\psi} - v - \hat{\psi}).$$

Using Lemma 1 it yields that, by a translation with $\hat{\psi}$, the problem (4) is equivalent to problem (1). Then we conclude that $y_0 + \hat{\psi} = y_\psi$.

Since $\int_\Omega \nabla y \nabla \hat{\psi} = -\int_\Omega \Delta \hat{\psi} y$ and $\Delta \hat{\psi} \in H^{-1}(\Omega)$, then we can consider the approximate problem

$$\min \left\{ \frac{1}{2} \int_\Omega |\nabla y|^2 - \int_\Omega \left(f + \Delta \hat{\psi} \right) y \ : \ y \in C_k \right\}, \tag{6}$$

where $C_k = \{y \in H_0^1(\Omega) : y(x_i) \geq 0, \forall i = 1, 2, \ldots, k\}$ and $\{x_i\}_i$ is a dense set in Ω.

In (6), we assume that $dim\ \Omega = 1$ (for the numerical applications in Sect. 3), but the result can be extended in higher dimension by using non hilbertian Sobolev spaces. Using the Sobolev imbedding theorem and the weak lower semicontinuity of the norm, then we can prove the following approximation result (see Merluşcă [9])

Theorem 1. *The sequence* $\{\bar{y}_k\}_k$ *of the solutions of problems* (6)*, for* $k \in \mathbb{N}$*, is a strongly convergent sequence in* $H_0^1(\Omega)$ *to the unique solution* $\bar{y}$ *of the problem* (4)*.*

We denote $\hat{f} = f + \Delta\hat{\psi} \in H^{-1}(\Omega)$. Applying the Fenchel duality theorem to problem (6) we obtain the dual problem

$$\min\left\{\frac{1}{2}|y^* + \hat{f}|^2_{H^{-1}(\Omega)} \quad : \; y \in C_k^*\right\}, \tag{7}$$

where $C_k^* = \{y^* \in H^{-1}(\Omega) : y^* = \sum_{i=1}^k \alpha_i \delta_{x_i}, \alpha_i \geq 0\}$ is the dual cone.

Remark 2. Let y_k^* be the solution of the dual approximate problem (7). Since $y_k^* \in C_k^*$, it is sufficient to compute the coefficients α_i^*, due to the formula

$$y_k^* = \sum_{i=1}^k \alpha_i^* \delta_{x_i}.$$

The solution y_k of the approximate problem (6) is computed using the equality $y_k = J^{-1}(y_k^* + \hat{f})$, where J is the duality mapping $J : H_0^1(\Omega) \to H^{-1}(\Omega)$ and we also have $\alpha_i^* y_k(x_i) = 0, \quad \forall i = \overline{1,k}$.

We obtain the formula for the solution of the approximate problem, denoted by y_k^0,

$$y_k^0 = \sum_{i=1}^k \alpha_i^* J^{-1}(\delta_i) + J^{-1}(\hat{f})$$

using the fact that the duality mapping $J : H_0^1(\Omega) \to H^{-1}(\Omega)$ is defined by $J(y) = -\Delta y$.

Then applying (5) we find the approximate solution of the general obstacle problem (1).

3 Numerical Applications

In this section we discuss two examples in one dimension for the general obstacle problem for second order operators.

We consider the obstacle problem (1) with $\Omega =]-1, 1[$ the domain, $\psi \equiv -1/18$ the obstacle and

$$f(x) = \begin{cases} -1, & |x| > 1/4\,, \\ 1 - 32x^2, & |x| \leq 1/4. \end{cases}$$

The solution of this problem is, Ockendon and Elliott [13], (pp. 93–94)

$$u(x) = \begin{cases} -\frac{1}{18} + \frac{1}{2}\left(x \pm \frac{2}{3}\right)^2, & \frac{2}{3} < |x| \leq 1, \\ -\frac{1}{18}, & \frac{1}{3} \leq |x| \leq \frac{2}{3}, \\ -\frac{1}{18} + \frac{1}{2}\left(x \pm \frac{1}{3}\right)^2, & \frac{1}{4} \leq |x| < \frac{1}{3}, \\ -\frac{1}{32} + \frac{8}{3}x^2\left(x^2 - \frac{3}{16}\right), & |x| < \frac{1}{4}. \end{cases}$$

The duality mapping $J : H_0^1(\Omega) \to H^{-1}(\Omega)$ is defined as $J(y) = -y''$. It is a linear bounded operator.

We consider the discretization $x_i = 2ih - 1$, where $h = 1/k$, for all $i \in \{0, 1, 2, \ldots, k\}$.

We denote by $d_i = J^{-1}(\delta_{x_i})$. Computing them we get

$$d_i(x) = \begin{cases} 0.5(1 - x_i)(x + 1), & x \le x_i, \\ 0.5(1 - x_i)(x + 1) - x + x_i, & x > x_i, \end{cases} \quad \forall i \in \{0, 1, \ldots, k\}.$$

We consider $y_f = J^{-1}(f + \Delta\hat{\psi})$ the solution of the problem

$$\begin{aligned} -y_f'' &= f + \Delta\hat{\psi}, \qquad \text{on }]-1, 1[, \\ y_f(-1) &= y_f(1) = 0. \end{aligned}$$

Then

$$|y^* + f + \Delta\hat{\psi}|^2_{H^{-1}(\Omega)} = \left| \sum_{i=1}^{k} \alpha_i d_i + y_f \right|^2_{H_0^1(\Omega)}$$

and

$$\left| \sum_{i=1}^{k} \alpha_i d_i + y_f \right|^2_{H_0^1(\Omega)} = \sum_{i,j=1}^{k} \alpha_i \alpha_j \int_\Omega d_i' d_j' + 2 \sum_{i=1}^{k} \alpha_i \int_\Omega d_i' y_f' + \int_\Omega (y_f')^2.$$

Denoting $a_{ij} = \int_\Omega d_i' d_j'$, $b_i = \int_\Omega d_i' y_f'$, we solve the dual problem (7) which is equivalent to the quadratic minimization problem

$$\min_{\alpha \ge 0} \frac{1}{2} \alpha^T A \alpha + b^T \alpha, \tag{8}$$

where $A = [a_{ij}]$ and $b = [b_i]$.

Computing the components, we get $b_i = y_f(x_i)$ and

$$a_{ij} = \begin{cases} 0.5(1 + x_i)(1 - x_j), & j > i, \\ 0.5(1 + x_j)(1 - x_i), & j \le i. \end{cases}$$

In Fig. 1 we represent the coefficients $\{\alpha_i^*\}_{i=1,100}$, the solution of the problem (8).

We construct the solution of the problem (6) using Remark 2, then we apply formula (5) to obtain the approximate solution of (1).

In Fig. 2 we represent three solutions: the one computed by the duality method, the one computed with the `IPOPT` optimizer (`Freefem++` script included in the `ff-Ipopt` dynamic library, Hatch [7]; for details about the method see Wächter and Biegler [16]) and the exact solution given by Ockendon and Elliot [13]. They coincide graphically.

We now consider an example with general obstacle:

$$\min \left\{ \frac{1}{2} \int_\Omega |\nabla y|^2 - \int_\Omega f y \ : \ y \in K_\psi \right\}, \tag{9}$$

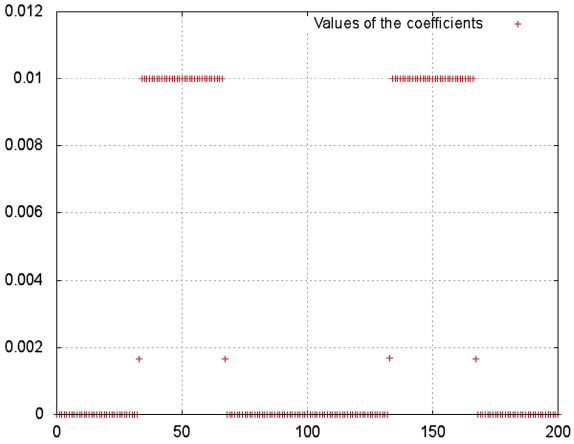


Fig. 1. The coefficients $\{\alpha_i^*\}_i$.

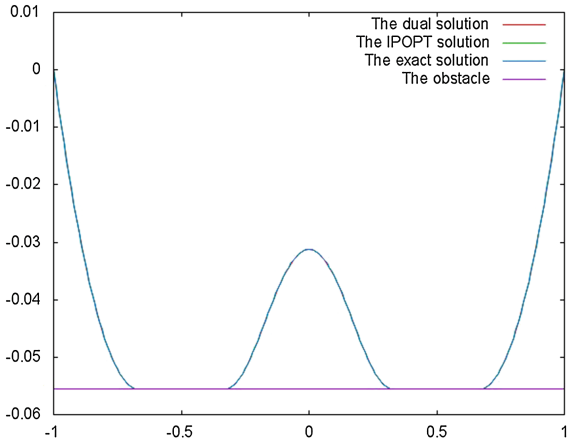


Fig. 2. The exact solution, the dual solution and the `IPOPT` solution are graphically identical.

where $K_\psi = \{y \in H_0^1(\Omega) : y \geq \psi\}$, $\Omega =]-1, 1[$, $\psi(x) = -x^2 + 0.5$ and

$$f(x) = \begin{cases} -10, & |x| > 1/4, \\ 10 - x^2, & |x| \leq 1/4. \end{cases}$$

After solving the quadratic minimization problem (8), the solution of which is represented in Fig. 3, we compute the solution of the approximate problem (6) using the Remark 2.

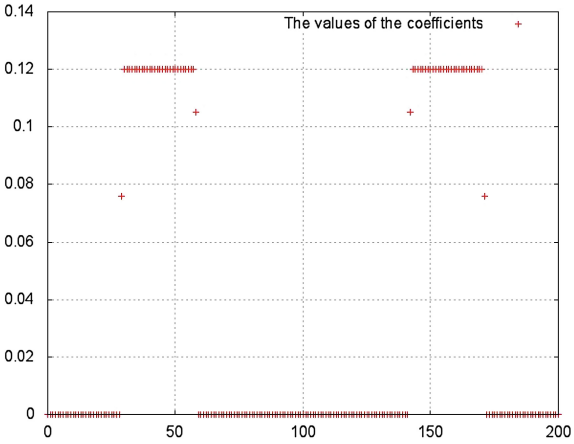


Fig. 3. The coefficients $\{\alpha_i^*\}_i$.

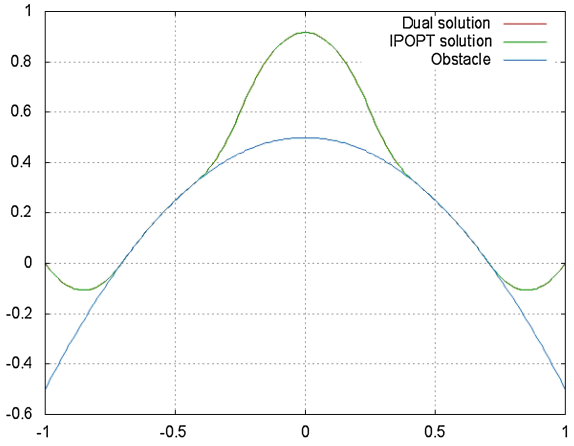


Fig. 4. The dual solution and the `IPOPT` solution coincide graphically.

We represent in Fig. 4 the obstacle ψ and the solutions, one computed by the duality method and the other one computed by the `IPOPT` method [7]. The two solutions are very close and coincide graphically.

References

1. Barbu, V.: Optimal Control of Variational Inequalities. Research Notes in Mathematics, vol. 100. Pitman, Boston (1984)
2. Barbu, V., Precupanu, T.: Convexity and Optimization in Banach Spaces. Noordhoff, London (1978)
3. Ciarlet, P.G.: Numerical analysis of the finite element method, Les Presses de l'Universitè de Montrèal (1976)

4. Caffarelli, L.A., Friedman, A.: The obstacle problem for the biharmonic operator. Annali della Scuola Normale Superiore di Pisa-Classe di Scienze **6**(1), 151–184 (1979)
5. Duvaut, G., Lions, J.-L.: Les inéquations en mécanique et en physique. Travaux et Recherches Mathematiques, N. 21. Dunod, Paris (1972)
6. Glowinski, R.: Numerical Methods for Nonlinear Variational Problems. Springer, New York (1984)
7. Hatch, F.: Freefem documetation, Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, Paris (2013). http://www.freefem.org/ff++/ftp/freefem++doc.pdf
8. Merluşcă, D.R.: A duality algorithm for the obstacle problem. Ann. Acad. Rom. Sci. **5**(1–2), 209–215 (2013)
9. Merluşcă, D.R.: A duality-type method for the obstacle problem. Analele Stiintifice ale Universitatii Ovidius Constanta Seria Matematica **21**(3), 181–195 (2013)
10. Merluşcă, D.R.: A duality-type method for the fourth order obstacle problem, submitted to U.P.B. Sci. Bull. Ser. A.
11. Murea, C.M., Tiba, D.: A direct algorithm in some free boundary problems, BCAM Publications (2012). http://www.bcamath.org/documentos_public/archivos/publicaciones/obstacle_2012-07-06.pdf
12. Neittaanmaki, P., Sprekels, J., Tiba, D.: Optimization of Elliptic Systems: Theory and Applications. Springer Monographs in Mathematics. Springer, New York (2006)
13. Ockendon, M.C., Elliot, J.R.: Weak and Variational Methods for Moving Boundary Problems. Pitman, London (1982)
14. Rodrigues, J.F.: Obstacle Problems in Mathematical Physics. Elsevier Science, North-Holland (1987)
15. Sprekels, J., Tiba, D.: Sur les arches lipschitziennes C.R.A.S. Paris Ser. I Math. **331**(2), 179–184 (2000). (French)
16. Wächter, A., Biegler, L.T.: On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. Math. Program. **106**(1), 25–57 (2006)

A Penalization Method for the Elliptic Bilateral Obstacle Problem

Cornel Marius Murea[1(✉)] and Dan Tiba[2,3]

[1] Laboratoire de Mathématiques, Informatique et Applications, Université de Haute Alsace, 4–6, Rue des Frères Lumière, 68093 Mulhouse Cedex, France
cornel.murea@uha.fr
http://www.edp.lmia.uha.fr/murea/
[2] Institute of Mathematics of the Romanian Academy, P.O. Box 1–764, 014700 Bucharest, Romania
[3] Academy of Romanian Scientists, Bucharest, Romania
dan.tiba@imar.ro

Abstract. In this paper we propose a new algorithm for the wellknown elliptic bilateral obstacle problem. Our approach enters the category of fixed domain methods and solves just linear elliptic equations at each iteration. The approximating coincidence set is explicitly computed. In the numerical examples, the algorithm has a fast convergence.

Keywords: Obstacle problem · Free boundary problems · Penalization

1 Introduction

The obstacle problem may be formulated as an elliptic variational inequality. Detailed theoretical discussions of various variational inequalities may be found in [4,15,23]. Applications, including optimal control problems are investigated in the books [1,5,6,24]. From the point of view of the numerical approximation, we quote just the monographs [6,7,21].

In this paper we propose an algorithm for the elliptic bilateral obstacle problem which is of fixed domain type in the sense that the finite element discretization is given in the whole domain, independently of the position of the unknown free boundary. In each iteration a linear elliptic equation has to be solved in the whole domain and the corresponding stiffness matrix is common for all iterations. This is a clear advantage from the point of view of the implementation and the approximating coincidence set is explicitly computed in each iteration and it converges in the Hausdorff-Pompeiu sense [20] to the searched geometry. Moreover, we need just a scalar penalization parameter in our method. A similar strategy was employed in [18] for the elliptic unilateral obstacle problem and for parabolic variational inequalities.

Dan Tiba—Supported by Grant 145/2011 CNCS, Romania.

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 189–198, 2014.
DOI: 10.1007/978-3-662-45504-3_18

Our approach is inspired from shape optimization techniques, but no shape optimization problem is used here although this is a known method in free boundary problems, [2]. One may compare the present approach to the recent works [8,19,22]. An efficient Lagrangian method together with a primal-dual active set strategy with regularization is studied in [13]. But our approach and arguments are certainly diferent. We also quote the multi grid method employed in [12], the path-following method for semi-smooth Newton schemes [10] and a duality-type method [16,17].

2 Formulation of the Problem and the Algorithm

Let D be a smooth domain in $\mathbb{R}^d$, $d \in \mathbb{N}^*$ and $f \in L^2(D)$ be given. We denote the obstacles by $\psi_1, \psi_2 : D \to \mathbb{R}$, such that $\psi_1, \psi_2 \in H^2(D)$, $\psi_1 \leq \psi_2$ in D, $\psi_{1|\partial D} \leq 0$, $\psi_{2|\partial D} \geq 0$. The admisible set

$$K = \left\{ v \in H_0^1(D);\ \psi_1(x) \leq v(x) \leq \psi_2(x) \text{ a.e. in } D \right\}$$

is a nonvoid closed convex subset of $H_0^1(D)$.

To K, the following variational inequality, may be associated:

$$\int_D \nabla y \cdot (\nabla y - \nabla v)\, dx \leq \int_D f\,(y - v)\, dx, \quad \forall v \in K. \tag{1}$$

The existence of a unique solution $y \in K$ is wellknown.

We introduce $\beta \subset \mathbb{R} \times \mathbb{R}$ the maximal monotone graph given by

$$\beta(r) = \begin{cases} \emptyset, & r < 0, \\ ]-\infty, 0], & r = 0, \\ 0, & r > 0, \end{cases} \tag{2}$$

the maximal monotone graph $\gamma \subset \mathbb{R} \times \mathbb{R}$ given by

$$\gamma(r) = \begin{cases} 0, & r < 0, \\ [0, +\infty[, & r = 0, \\ \emptyset, & r > 0 \end{cases} \tag{3}$$

and denote by β_ϵ, γ_ϵ, $\epsilon > 0$, their Yosida approximations. We have

$$\beta_\epsilon(r) = \begin{cases} \frac{1}{\epsilon} r, & r \leq 0, \\ 0, & r > 0, \end{cases} \qquad \gamma_\epsilon(r) = \begin{cases} 0, & r < 0, \\ \frac{1}{\epsilon} r, & r \geq 0. \end{cases}$$

Notice that β_ϵ is a concave and γ_ϵ is a convex function in $\mathbb{R}$.

In the case when $f \in L^2(D)$, $\psi_1, \psi_2 \in H^2(D)$ with the compatibility condition $\psi_1 \leq \psi_2$ in D, $\psi_{1|\partial D} \leq 0$, $\psi_{2|\partial D} \geq 0$, it is known that the solution of (1) satisfies the regularity property $y \in H^2(D)$. Moreover, in this case, the obstacle problem may be written as a multivalued equation

$$-\Delta y + \beta(y - \psi_1) + \gamma(y - \psi_2) \ni f \text{ in } D. \tag{4}$$

One can define two coincidence sets, corresponding to the two obstacles:

$$D_1 = \{x \in D;\ y(x) = \psi_1(x)\}$$
$$D_2 = \{x \in D;\ y(x) = \psi_2(x)\}$$

and associated to (1).

We state now our algorithm.

Algorithm

(1) Choose $n = 0$, $\epsilon_0 > 0$, $\Omega_1^0 \subset D$, $\Omega_2^0 \subset D$ open subsets such that $(D \setminus \Omega_1^0) \cap (D \setminus \Omega_2^0) = \emptyset$, $\tilde{y}_{-1} = 0$;
(2) Compute $y_n \in H_0^1(D)$ as solution of the linear elliptic equation

$$-\Delta y_n + \frac{1}{\epsilon_n}\chi_{D\setminus\Omega_1^n}(y_n - \psi_1) + \frac{1}{\epsilon_n}\chi_{D\setminus\Omega_2^n}(y_n - \psi_2) = f \text{ in } D \tag{5}$$

(3) Compute $\mathbf{y}_n = \min\{\psi_2,\ \max\{y_n, \psi_1\}\}$, $\Omega_1^{n+1} = \{x \in D;\ \mathbf{y}_n(x) > \psi_1(x)\}$, $\Omega_2^{n+1} = \{x \in D;\ \mathbf{y}_n(x) < \psi_2(x)\}$ $\epsilon_{n+1} = \frac{\epsilon_n}{2}$;
(4) If $\|\mathbf{y}_n - \mathbf{y}_{n-1}\|_{H^1(D)} < tol$ then STOP else n=n+1 GO TO step 2.

Remark 1. By the classical result of [3], the elastic-plastic torsion problem is equivalent with a variational inequality of obstacle type and our algorithm may be applied as well.

We convene to extend the value $\frac{1}{\epsilon}$ for β'_ϵ and γ'_ϵ in the origin and we can rewrite the step 2 of the Algorithm as

$$-\Delta y_n + \left(\beta'_{\epsilon_n}(y_{n-1} - \psi_1)\right)(y_n - \psi_1) + \left(\gamma'_{\epsilon_n}(y_{n-1} - \psi_2)\right)(y_n - \psi_2) = f. \tag{6}$$

Recall that the usual approximation by regularization of the variational inequality (1) is

$$-\Delta\tilde{y}_n + \beta_{\epsilon_n}(\tilde{y}_n - \psi_1) + \gamma_{\epsilon_n}(\tilde{y}_n - \psi_2) = f \text{ in } D, \tag{7}$$

plus homogeneous boundary conditions on ∂D.

Notice that $\beta_\epsilon(r) = \beta'_\epsilon(r)r$ and $\gamma_\epsilon(r) = \gamma'_\epsilon(r)r$, under the above convention, which shows that (6) and (7) have very similar structure. Clearly, (7) is a nonlinear elliptic equation, while the decoupling operated in (6) allows to use linear elliptic equations.

3 Stability

We present in this section a stability result in $L^2(D)$ for the algorithm introduced above applied to the bilateral obstacle problem.

Theorem 1. *i) The sequence $\{y_n\}$ is bounded in $L^2(D)$.*

ii) There is $C > 0$, independent of n, such that:

$$\int_{D\setminus\Omega_2^n}(y_n-\psi_2)_+^2dx+\int_{D\setminus\Omega_1^n}(y_n-\psi_1)_-^2dx\le C\epsilon_n. \tag{8}$$

Proof. Using $\beta_\epsilon(r)=\beta'_\epsilon(r)r$, the concavity of $\beta_\epsilon(\cdot)$ and the definition of the subdifferential of concave mapping, we obtain

$$\begin{aligned}\left(\beta'_{\epsilon_n}(y_{n-1}-\psi)\right)(y_n-\psi)=\left(\beta'_{\epsilon_n}(y_{n-1}-\psi)\right)(y_{n-1}-\psi)\\ +\left(\beta'_{\epsilon_n}(y_{n-1}-\psi)\right)(y_n-\psi-y_{n-1}+\psi)\ge\beta_{\epsilon_n}(y_{n-1}-\psi)\\ +\beta_{\epsilon_n}(y_n-\psi)-\beta_{\epsilon_n}(y_{n-1}-\psi)=\beta_{\epsilon_n}(y_n-\psi).\end{aligned}$$

We use the above inequality in the Eq. (5). We get

$$-\Delta y_n+\beta_{\epsilon_n}(y_n-\psi_1)+\frac{1}{\epsilon_n}\chi_{D\setminus\Omega_2^n}(y_n-\psi_2)\le f, \tag{9}$$

where β is given by (2) and β_{ϵ_n} is its regularization.

We multiply (9) by $(y_n-\psi_2)_+$ and we use that

$$\beta_{\epsilon_n}(y_n-\psi_1)(y_n-\psi_2)_+=0. \tag{10}$$

While $\beta_{\epsilon_n}(y_n-\psi_1)$ may take negative values, this happens for $y_n\le\psi_1$, that is $y_n-\psi_2\le 0$ (since $\psi_1\le\psi_2$). Then $(y_n-\psi_2)_+=0$ and (10) follows. We infer

$$\begin{aligned}&\int_D|\nabla(y_n-\psi_2)_+|^2+\frac{1}{\epsilon_n}\int_{D\setminus\Omega_2^n}(y_n-\psi_2)_+^2 \\ &\le\int_D f(y_n-\psi_2)_++\int_D\nabla\psi_2\cdot\nabla(y_n-\psi_2)_+.\end{aligned} \tag{11}$$

By the conditions $\psi_{2\,|\partial D}\ge 0$ and $y_{n\,|\partial D}=0$, we have $(y_n-\psi_2)_+=0$ on ∂D and the Poincaré inequality shows that $\{(y_n-\psi_2)_+\}$ is bounded in $H_0^1(D)$, by (11).

Equation (5) may be rewritten in the form

$$-\Delta y_n+\frac{1}{\epsilon_n}\chi_{D\setminus\Omega_1^n}(y_n-\psi_1)+\gamma'_{\epsilon_n}(y_{n-1}-\psi_2)(y_n-\psi_2)=f. \tag{12}$$

We compute

$$\begin{aligned}\gamma'_{\epsilon_n}(y_{n-1}-\psi_2)(y_n-\psi_2)=\gamma'_{\epsilon_n}(y_{n-1}-\psi_2)(y_{n-1}-\psi_2)\\ +\gamma'_{\epsilon_n}(y_{n-1}-\psi_2)(y_n-\psi_2-y_{n-1}+\psi_2)=\gamma_{\epsilon_n}(y_{n-1}-\psi_2)\\ +\gamma'_{\epsilon_n}(y_{n-1}-\psi_2)(y_n-\psi_2-y_{n-1}+\psi_2)\le\gamma_{\epsilon_n}(y_{n-1}-\psi_2)\\ +\gamma_{\epsilon_n}(y_n-\psi_2)-\gamma_{\epsilon_n}(y_{n-1}-\psi_2)=\gamma_{\epsilon_n}(y_n-\psi_2)\end{aligned} \tag{13}$$

using the subdifferential of convex mappings.

By (12), (13), we obtain

$$-\Delta y_n + \frac{1}{\epsilon_n}\chi_{D\setminus\Omega_1^n}(y_n-\psi_1) + \gamma_{\epsilon_n}(y_n-\psi_2) \geq f. \tag{14}$$

Multiply (14) by $-(y_n-\psi_1)_- \in H_0^1(D)$, due to $y_n = 0$ on ∂D, $\psi_1 \leq 0$ on ∂D:

$$\int_D |\nabla(y_n-\psi_1)_-|^2 + \frac{1}{\epsilon_n}\int_{D\setminus\Omega_1^n}(y_n-\psi_1)_-^2 - \gamma_{\epsilon_n}(y_n-\psi_2)(y_n-\psi_1)_- \tag{15}$$
$$\leq -\int_D f(y_n-\psi_1)_- + \int_D \nabla\psi_1\cdot\nabla(y_n-\psi_1)_-.$$

Notice that

$$-\gamma_{\epsilon_n}(y_n-\psi_2)(y_n-\psi_1)_- = 0 \tag{16}$$

since $-\gamma_{\epsilon_n}(y_n-\psi_2)$ may take negative values just for $y_n \geq \psi_2 \geq \psi_1$ and in this case $(y_n-\psi_1)_- = 0$. By (15), (16) we obtain:

$$\int_D |\nabla(y_n-\psi_1)_-|^2 + \frac{1}{\epsilon_n}\int_{D\setminus\Omega_1^n}(y_n-\psi_1)_-^2 \tag{17}$$
$$\leq -\int_D f(y_n-\psi_1)_- + \int_D \nabla\psi_1\cdot\nabla(y_n-\psi_1)_-.$$

Relation (17) shows that $\{(y_n-\psi_1)_-\}$ is bounded in $H_0^1(D)$, by the Poincaré inequality.

We use the inequality

$$(x-b)_+ \leq (x-a)_+ + (a-b)_+$$

and we have

$$(y_n-\psi_1)_+ \leq (y_n-\psi_2)_+ + (\psi_2-\psi_1)_+ = (y_n-\psi_2)_+ + \psi_2 - \psi_1.$$

Relation (11) shows that $\{(y_n-\psi_1)_+\}$ is bounded in $L^2(D)$. In combination with (17), it yields $\{y_n\}$ bounded in $L^2(D)$.

Relation (8) follows by adding (11) and (17) and using the already established boundedness of all the terms except the penalization term. This ends the proof.

Remark 2. In fact, the above proof shows that $\{y_n\}$ bounded in $L^p(D)$, $p > 2$ depending on the dimension of D. Relation (8) says that the sequence $\{y_n\}$ does not overpass the obstacles ψ_1, ψ_2, in the limit. The proof also provides partial information on $\{\nabla y_n\}$, but it is unclear whether $\{y_n\}$ is bounded in $H_0^1(D)$.

4 Numerical Tests

We have used the software FreeFem++ v 3.19, [9]. For all tests, we use the same initial guess for the coincidence set $D\setminus\Omega_1^0 = \emptyset$ and $D\setminus\Omega_2^0 = \emptyset$.

Test 1. We consider the torsion of an elastic-plastic prism studied in [7] p. 133 and [25]. The cross-section of the prism is $D = [0,1]\times[0,1]$. We solve the problem (1) with $f(x) = -8$ where $K = \{v \in H_0^1(D);\ -1 \leq \nabla v(x) \leq 1 \text{ a.e. in } D\}$. For $v \in K$, the set $\{x \in D;\ |\nabla v(x)| < 1\}$ is the elastic zone and $\{x \in D;\ |\nabla v(x)| = 1\}$ is the plastic zone, [5, p. 264]. By the result of [3], the elastic-plastic torsion problem is equivalent with a variational inequality of obstacle type $K = \{v \in H_0^1(D);\ \psi_1(x) \leq v(x) \leq \psi_2(x) \text{ a.e. in } D\}$ where $\psi_1(x) = -dist(x, \partial D)$ and $\psi_2(x) = dist(x, \partial D)$. If $f < 0$, then $y \leq 0$, consequently the top obstacle will be inactive.

We use a mesh of 39158 triangles, 19836 vertices and size $h = \frac{1}{128}$. The tolerance for the stopping test is $tol = 10^{-3}$ and the penalization parameter is $\epsilon_n = 0.003$. The coincidence set of the solution presented in Fig. 1 at the right is similar to the above references.

Fig. 1. Test 1. The computed coincidence set of the plastic zone for the bottom obstacle at the first (left), second (middle) and last (right) iteration.

Our algorithm stops after 6 iterations and the relative error in the H^1 norm at the last iteration is $\|\mathbf{y}_n - \mathbf{y}_{n-1}\|_{H^1(D)} = 1.5 \times 10^{-5}$.

In [18], we have tested numerically with positive results the stability of a similar algorithm when f the right-hand side in (1) is perturbed.

Test 2. We solve the problem (1) where

$$K = \{v \in H_0^1(D);\ \psi_1(x) \leq v(x) \leq \psi_2(x) \text{ a.e. in } D\},$$

$D = [0,1] \times [0,1]$, $\psi_1(x) = -dist(x, \partial D)$, $\psi_2(x) = dist(x, \partial D)$ and $f(x) = 11(x + y - 1)$. Now both obstacles are active.

We use a mesh of 39158 triangles, 19836 vertices, the size $h = \frac{1}{128}$, the tolerance for the stopping test $tol = 10^{-3}$ and the penalization parameter is $\epsilon_n = 0.003$. The algorithm stops in 4 iterations and the relative error in the H^1 norm at the last iteration is $\|\mathbf{y}_n - \mathbf{y}_{n-1}\|_{H^1(D)} = 0.000209$.

The coincidence sets are presented in Fig. 2 and the computed solution in Fig. 3.

We solved the problem on different meshes, see Table 1. We denote by $\mathbf{u}_i$, the solution obtained using the mesh no. i.

In [11], for the semi-smooth Newton method, it is proved a mesh-independence result: the continuous and the discrete process, converge q-linearly with the same rate.

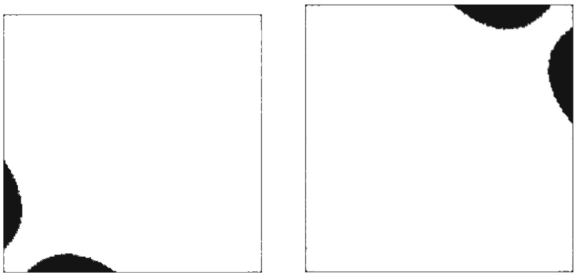


Fig. 2. Test 2. Coincidence sets of the plastic zone for the bottom obstacle (left) and for the top obstacle (right).

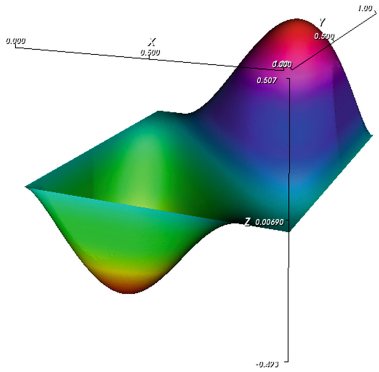


Fig. 3. Test 2. Computed solution.

Table 1. Test 2. Mesh parameters.

Mesh no.	Mesh size h	Triangles	Vertices	$\|\mathbf{u}_i - \mathbf{u}_{i-1}\|_{H^1(D)}$
1	1/64	9720	4989	-
2	1/128	39158	19836	0.019005
3	1/256	154050	77538	0.008826
4	1/512	630326	316188	0.005768

Test 3. Now we test the algorithm for the torsion of the elastic-plastic prism discussed in [14,26]. We can put this problem in the form (1). Let $D = [0,1] \times [0,1]$, $\psi_1(x,y) = -dist\,((x,y),\partial D)$, $\psi_2(x,y) = 0.2$ for all $(x,y) \in D$ and set

$$g(x) = \begin{cases} 6x, & 0 < x \le 1/6, \\ 2(1-3x), & 1/6 < x \le 1/3, \\ 6(x-1/3), & 1/3 < x \le 1/2, \\ 2\left(1-3(x-1/3)\right), & 1/2 < x \le 2/3, \\ 6(x-2/3), & 2/3 < x \le 5/6, \\ 2\left(1-3(x-2/3)\right), & 5/6 < x \le 1 \end{cases}$$

and

$$f(x,y)=\begin{cases} 300, & (x,y)\in S=\{(x,y)\in D;|x-y|\le 0.1\ \&\ x\le 0.3\},\\ -70\exp(y)g(x), & x\le 1-y \text{ and } (x,y)\notin S,\\ 15\exp(y)g(x), & x>1-y \text{ and } (x,y)\notin S.\end{cases}$$

We use a mesh of 39158 triangles, 19836 vertices, the size $h=\frac{1}{128}$, the tolerance for the stopping test $tol=10^{-3}$ and the penalization parameter is $\epsilon_n=0.03$. The computed solution after 6 iterations is presented in Fig. 4 and the corresponding coincidence sets in Fig. 5. The relative error in the H^1 norm at the last iteration is $\|\mathbf{y}_n-\mathbf{y}_{n-1}\|_{H^1(D)}=2.7\times 10^{-5}$.

In [14], an augmented lagrangian active set strategy is employed. At each iteration, a reduced linear system associated with the inactive set is solved. In [26], at each iteration, linear systems associated to the complementary of the coincidence sets are solved.

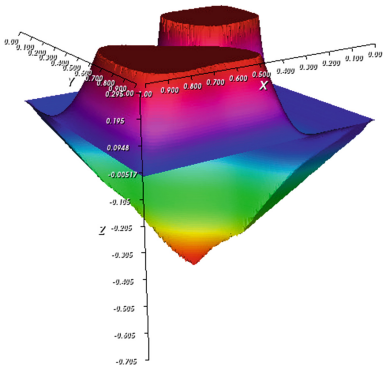


Fig. 4. Test 3. Computed solution.

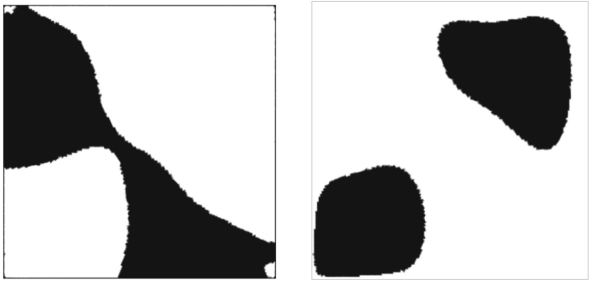


Fig. 5. Test 3. Coincidence sets for the bottom obstacle (left) and for the top obstacle (right).

References

1. Barbu, V.: Optimal control of variational inequalities. Research Notes in Mathematics, vol. 100. Pitman, Boston (1984)
2. Burger, M., Matevosyan, N., Wolfram, M.T.: A level set based shape optimization method for an elliptic obstacle problem. Math. Models Methods Appl. Sci. **21**(4), 619–649 (2011)
3. Brézis, H., Sibony, M.: Équivalence de deux inéquations variationnelles et applications. (French). Arch. Rational Mech. Anal. **41**, 254–265 (1971)
4. Brézis, H.: Problèmes unilatéraux. J. Math. Pures Appl. **9**(51), 1–168 (1972)
5. Duvaut, G., Lions, J.-L.: Les inéquations en mécanique et en physique. Travaux et Recherches Mathématiques, vol. 21. Dunod, Paris (1972)
6. Elliott, C.M., Ockendon, J.R.: Weak and variational methods for moving boundary problems. Research Notes in Mathematics, vol. 59. Pitman, London (1982)
7. Glowinski, R., Lions, J.-L., Trémolières, R.: Numerical analysis of variational inequalities. Studies in Mathematics and its Applications, vol. 8. North-Holland Publishing Co., Amsterdam-New York (1981)
8. Halanay, A., Murea, C.M., Tiba, D.: Existence and approximation for a steady fluid-structure interaction problem using fictitious domain approach with penalization. Mathematics and its Applications **5**(1–2), 120–147 (2013)
9. Hecht, F.: http://www.freefem.org
10. Hintermuller, M., Kunisch, K.: Path-following methods for a class of constrained minimization problems in function space. SIAM J. Optim. **17**(1), 159–187 (2006)
11. Hintermuller, M., Ulbrich, M.: Math. Program. Ser. B. A mesh-independence result for semismooth Newton methods **101**, 151–184 (2004)
12. Hoppe, R.H.W., Kornhuber, R.: Adaptive multilevel methods for obstacle problems. SIAM J. Numer. Anal. **31**(2), 301–323 (1994)
13. Ito, K., Kunisch, K.: Semi-smooth Newton methods for variational inequalities of the first kind. M2AN Math. Model. Numer. Anal. **37**(1), 41–62 (2003)
14. Karkkainen, T., Kunisch, K., Tarvainen, P.: Augmented Lagrangian active set methods for obstacle problems. J. Optim. Theory Appl. **119**(3), 499–533 (2003)
15. Kinderlehrer, D., Stampacchia, G.: An introduction to variational inequalities and their applications. Classics in Applied Mathematics, vol. 31. SIAM, Philadelphia (2000). Reprint of the 1980 original
16. Merlusca, D.: A duality algorithm for the obstacle problem. Ann. Acad. Rom. Sci. **5**(1–2), 209–215 (2013)
17. Merlusca, D.: A duality-type method for the fourth order obstacle problem. U.P.B. Sci. Bull. Ser. A **76**(2), 147–158 (2014)
18. Murea, C.M., Tiba, D.: A direct algorithm in some free boundary problems, Submitted to J. Numer. Math. (2014)
19. Neittaanmaki, P., Pennanen, A., Tiba, D.: Fixed domain approaches in shape optimization problems with Dirichlet boundary conditions. Inverse Prob. **25**(5), 1–18 (2009)
20. Neittaanmaki, P., Sprekels, J., Tiba, D.: Optimization of Elliptic Systems. Theory and Applications. Springer Monographs in Mathematics. Springer, New York (2006)
21. 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 Inc., New York (1994)

22. Neittaanmaki, P., Tiba, D.: Fixed domain approaches in shape optimization problems. Inverse Prob. **28**(9), 1–35 (2012)
23. Rodrigues, J.-F.: Obstacle Problems in Mathematical Physics. North-Holland Publishing Co., Amsterdam (1987)
24. Tiba, D.: Optimal Control of Nonsmooth Distributed Parameter Systems. Lecture Notes in Mathematics, vol. 1459. Sprringer, Berlin (1990)
25. Xue, L., Cheng, X.-L.: An algorithm for solving the obstacle problems. Comput. Math. Appl. **48**(10–11), 1651–1657 (2004)
26. Wang, F., Cheng, X.-L.: An algorithm for solving the double obstacle problems. Appl. Math. Comput. **201**(1–2), 221–228 (2008)

Binary Level Set Method for Topology Optimization of Variational Inequalities

Andrzej Myśliński(✉)

Systems Research Institute, ul. Newelska 6, 01-447 Warsaw, Poland
myslinsk@ibspan.waw.pl

Abstract. The paper is concerned with the 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 binary level set method is used to represent sub-domains rather than the standard method. The binary level set function takes at convergence value 1 in each sub domain of a whole design domain and -1 outside this sub domain. The sub domains interfaces are represented by discontinuities of these functions. Using a two-phase approximation and a binary level set approach the original structural optimization problem is reformulated as an equivalent constrained optimization problem in terms of this level set function. Necessary optimality condition is formulated. Numerical examples are provided and discussed.

Keywords: Topology optimization · Unilateral contact problems · Binary level set method · Uzawa method

1 Introduction

Topology optimization problem for an elliptic second order variational inequality is considered in the paper. This inequality governs unilateral contact between an elastic body and a rigid foundation. The results concerning the existence and the uniqueness of solutions to this inequality are provided in [9]. The topology optimization problem for the elastic body in unilateral contact consists in finding such material distribution within the domain occupied by the body in contact and/or 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.

Topology optimization of continuum structures is widely investigated in literature [1,5,7,11]. Among others the homogenization method and its simplified version solid isotropic material with penalization method (SIMP) as well as evolutionary structural optimization method have been proposed to solve these problems (see [5]). Recently, the level set approach [13] is employed in the numerical algorithms of structural optimization [7] for tracking the evolution of the domain boundary on a fixed mesh and finding an optimal solution to

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 199–209, 2014.
DOI: 10.1007/978-3-662-45504-3_19

structural optimization problems. This approach, in classical form, 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. The evolution of the domain boundary is governed by Hamilton-Jacobi equation. 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.

In order to avoid the drawbacks of the classical level set method an alternative piecewise constant level set method has been proposed, first in image processing area and next in structural optimization [14]. For a domain divided into 2^N sub-domains in classical level set approach is required 2^N level set functions to represent them. Piecewise constant level set method can identify an arbitrary number of sub-domains using only one discontinuous piecewise constant level set function. This function takes distinct constant values on each sub-domain. The interfaces between sub-domains are represented implicitly by the discontinuity of a set of characteristic functions of the sub-domains [14]. 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. Binary level set method [4,10,15] is a special piecewise constant level set method where the function takes only two values either $+1$ or -1. Compared with general piecewise constant level set approach binary level set approach requires N level set functions to represent a structure of 2^N different material phases and is very close to the phase-field approach [12].

In the paper the original structural optimization problem is approximated by a two-phase material optimization problem. Using the binary level set method this approximated problem is reformulated as an equivalent constrained optimization problem in terms of the binary level set function only. Therefore neither shape nor topological complicated sensitivity analysis is required. During the evolution of the binary level set function small holes can be created without the 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:

$$O_l = \{\Omega \subset R^2 : \Omega \text{ is open}, \quad E \subset \Omega \subset D, \quad {}^{\#}\Omega^c \leq l\}, \tag{1}$$

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

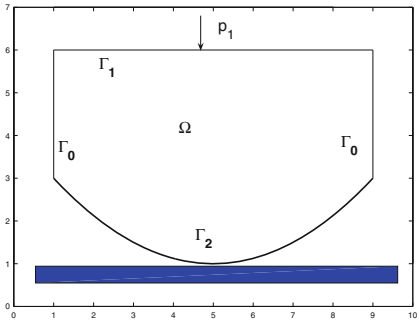


Fig. 1. Initial domain Ω.

$\delta\Omega$ 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 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

$$u_i(x) = 0 \quad \text{on} \quad \Gamma_0, \quad i = 1, 2, \tag{5}$$

$$\sigma_{ij}(x) n_j = p_i \quad \text{on} \quad \Gamma_1, \quad i, j = 1, 2, \tag{6}$$

$$u_N \leq 0, \quad \sigma_N \leq 0, \quad u_N \sigma_N = 0 \quad \text{on } \Gamma_2, \tag{7}$$

$$\mid \sigma_T \mid \leq 1, \quad u_T \sigma_T + \mid u_T \mid = 0 \quad \text{on } \Gamma_2, \tag{8}$$

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].

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*

$$\int_\Omega a_{ijkl} e_{ij}(u) e_{kl}(\varphi - u) dx - \int_\Omega f_i(\varphi_i - u_i) dx - \int_{\Gamma_1} p_i(\varphi_i - u_i) ds + \int_{\Gamma_2} \lambda(\varphi_T - u_T) ds \geq 0 \ \ \forall \varphi \in K, \tag{12}$$

$$\int_{\Gamma_2} (\zeta - \lambda) u_T ds \leq 0 \ \ \ \forall \zeta \in \Lambda, \tag{13}$$

$i, j, k, l = 1, 2$. Function λ is interpreted as a Lagrange multiplier corresponding to term $\mid u_T \mid$ in equality constraint in (8) [9]. 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]

$$Per(\Omega) \leq const_1, \ \ Per(\Omega) \stackrel{def}{=} \int_\Gamma dx. \tag{15}$$

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 [11] 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] the standard level set method [13] 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 [13]

$$\frac{\partial \phi(x,t)}{\partial t} + V(x,t) \cdot \nabla_x \phi(x,t) = 0. \tag{21}$$

Moreover $\phi(x,0) = \phi_0$ where $\phi_0(x)$ is a given function close to the signed distance function [13].

3.1 Binary Level Set Formulation

Recall from [4,10] the notion of a binary level set function. Let N be a given integer. Assume an open bounded domain D in R^2 is partitioned into 2^N sub-domains $\{\Omega_j\}_{j=1}^{2^N}$ such that

$$D = \bigcup_{j=1}^{2^N} (\Omega_j \cup \partial\Omega_j). \tag{22}$$

$\partial\Omega_j$ denotes the boundary of the sub-domain Ω_j. For $N = 2$ this function mapping $\phi : D \to R$, is defined as:

$$\phi(x) = \begin{cases} +1, \text{ if } x \in \Omega_1, \\ -1, \text{ if } x \in \Omega_2 = D \setminus \Omega_1. \end{cases} \tag{23}$$

The interface $\partial\Omega_1 = \partial\Omega_2$ is implicitly defined by the discontinuity of ϕ, i.e., $\partial\Omega_1 = \{x \in D \ : \ \phi(x) = \kappa, \quad \kappa \in (-1,1)\}$. In order to ensure that for every $x \in D$ this function converges to values $+1$ and -1 it is supposed to satisfy:

$$W(\phi) \stackrel{def}{=} (\phi^2 - 1) = 0. \tag{24}$$

More generally, using N binary level set functions ϕ_i, $i = 1, 2, ..., N$, satisfying (24) we can represent 2^N sub-domains Ω_j of D. The characteristic functions χ_j, $j = 1, 2, ..., 2^N$, of the sub-domains Ω_j in terms of binary level set functions ϕ_i are represented as

$$\chi_j = \frac{(-1)^{s(j)}}{2^N} \prod_{i=1}^{N} (\phi_i + 1 - 2b_i^{j-1}) \ \text{ where } \ s(j) = \sum_{i=1}^{N} b_i^{j-1}. \tag{25}$$

For $j = 1, 2, ..., 2^N$ and $i = 1, ..., N$ numbers $b_i^{j-1} = 0 \vee 1$ denotes binary representation of $j-1$th sub-domain. As long as each binary function satisfies (24) and $\chi_j(x)$ are defined by (25) then

$$\mathrm{supp}(\chi_\mathrm{j}) = \Omega_\mathrm{j}, \ \ \chi_\mathrm{j} = 1 \text{ in } \Omega_\mathrm{j} \quad \text{and} \quad \mathrm{supp}(\chi_\mathrm{i}) \cap \mathrm{supp}(\chi_\mathrm{j}) = \emptyset \ \text{ for } \ \mathrm{i} \neq \mathrm{j}, \tag{26}$$

$$\sum_{j} \mathrm{supp}(\chi_\mathrm{j}) = \Omega. \tag{27}$$

Condition (26) ensures non overlapping of the phases while (27) prevents vacuums. Basis functions χ_j are used to calculate the length of the boundary $\partial\Omega_j$ as well as the area inside Ω_j using the integrals:

$$\mid \partial\Omega_j \mid = \int_{\Omega_j} \mid \nabla\chi_j \mid dx \quad \text{and} \quad \mid \Omega_j \mid = \int_{\Omega_j} \chi_j dx. \tag{28}$$

The length of the boundary $\partial\Omega_j$ of sub-domain Ω_j equals the total variation of χ_j [2]. Consider piecewise constant density function $\rho = \rho(x) : D \to R^2$ defined as

$$\rho(x) = \begin{cases} c_1 \text{ if } x \in D \setminus \bar{\Omega}, \\ c_2 \text{ if } x \in \Omega, \end{cases} \tag{29}$$

where $0 < c_1 < c_2 < \infty$ denote two given material densities. This function can be constructed as a weighted sum of the characteristic functions χ_j. Denoting by $\{c_j\}_{j=1}^{2^N}$ a set of real scalars, we can represent a piecewise constant function ρ taking these 2^N distinct constant values in sub-domains Ω_j by

$$\rho(x) = \sum_{j=1}^{2^N} c_j \chi_j(x). \tag{30}$$

We confine to consider a two-phase problem in the domain D, i.e., we set $N = 2$ and $c_1 = \epsilon$, $\epsilon > 0$ as well as $c_2 = 1$. Since Ω consists from two sub-domains one binary level set function ϕ satisfying (24) will be used to describe these sub-domains. Therefore

$$\chi_1(x) = \phi(x) + 1 \ \text{ and } \ \chi_2(x) = 1 - \phi(x), \tag{31}$$

$$\rho(x) = \sum_{j=1}^{2} c_j \rho_j = c_1\chi_1(x) + c_2\chi_2(x) = c_1(\phi(x) + 1) + c_2(1 - \phi(x)). \tag{32}$$

Using (22) as well as (32) the structural optimization problem (19) can be transformed into the following one: *find* $\phi \in U_{ad}^{\phi}$ *such that*

$$\min_{\phi \in U_{ad}^{\phi}} J_\eta(\phi) = \int_{\Gamma_2} \rho(\phi)\sigma_N(u_\epsilon)\eta_N ds, \tag{33}$$

where the set U_{ad}^{ϕ} of the admissible functions is given as

$$U_{ad}^{\phi} = \{\phi \in H^1(D) : Vol(\phi) - Vol^{giv} \le 0,\ W(\phi) = 0,\ Per(\phi) \le const_1\}, \tag{34}$$

$$Vol(\phi) \stackrel{def}{=} \int_\Omega \rho(\phi) dx,\ Per(\phi) \stackrel{def}{=} \int_\Omega \mid \nabla\phi \mid dx. \tag{35}$$

The element $(u_\epsilon, \lambda_\epsilon) \in K \times \Lambda$ depending on ϵ satisfies the state system (12)–(13) in the domain D rather than Ω:

$$\int_D \rho(\phi) a_{ijkl} e_{ij}(u_\epsilon) e_{kl}(\varphi - u_\epsilon) dx - \int_D \rho(\phi) f_i(\varphi_i - u_{\epsilon i}) dx -$$
$$\int_{\Gamma_1} p_i(\varphi_i - u_{\epsilon i}) ds + \int_{\Gamma_2} \lambda_\epsilon(\varphi_T - u_{\epsilon T}) ds \geq 0 \ \ \forall \varphi \in K, \tag{36}$$

$$\int_{\Gamma_2} (\zeta - \lambda_\epsilon) u_{\epsilon T} ds \leq 0 \quad \forall \zeta \in \Lambda. \tag{37}$$

Lemma 2. *There exists an optimal solution $\phi \in H^1(D)$ to the optimization problem (33)–(37).*

The proof follows from the lower semicontinuity in $L^1(D)$ of the regularization term in (34) see [2, Theorem 3.2.1, p. 75].

3.2 Necessary Optimality Conditions

In order to formulate the necessary optimality condition for the optimization problem (33)–(37) we introduce the Lagrangian $L(\phi, \tilde{\lambda}) = L(\phi, u_\epsilon, \lambda_\epsilon, p^a, q^a, \tilde{\lambda})$:

$$L(\phi, \tilde{\lambda}) = J_\eta(\phi) + \int_D \rho(\phi) a_{ijkl} e_{ij}(u_\epsilon) e_{kl}(p^a) dx - \int_D \rho(\phi) f_i p_i^a dx$$
$$- \int_{\Gamma_1} p_i p_i^a ds + \int_{\Gamma_2} \lambda_\epsilon p_T^a ds + \int_{\Gamma_2} q^a u_{\epsilon T} ds + \tilde{\lambda} d(\phi) + \sum_{i=1}^{3} \frac{1}{2\mu_i} d_i^2(\phi), \tag{38}$$

where $i, j, k, l = 1, 2$, $\tilde{\lambda} = \{\tilde{\lambda}_i\}_{i=1}^3$, $d(\phi) = \{d_i(\phi)\}_{i=1}^3 = [Vol(\phi), W(\phi), Per(\phi)]^T$, $d^T(\phi)$ denotes a transpose of $d(\phi)$, $\mu_m > 0$, $m = 1, 2, 3$, is a given real. Element $(p^a, q^a) \in K_1 \times \Lambda_1$ denotes an adjoint state defined as follows:

$$\int_D \rho(\phi) a_{ijkl} e_{ij}(\eta + p^a) e_{kl}(\varphi) dx + \int_{\Gamma_2} q^a \varphi_T ds = 0 \quad \forall \varphi \in K_1, \tag{39}$$

$$\int_{\Gamma_2} \zeta(p_T^a + \eta_T) ds = 0 \ \ \forall \zeta \in \Lambda_1. \tag{40}$$

The sets K_1 and Λ_1 are given by

$$K_1 = \{\xi \in V_{sp} \ : \ \xi_N = 0 \ \text{ on } A^{st}\}, \tag{41}$$

$$\Lambda_1 = \{\zeta \in \Lambda \ : \ \zeta(x) = 0 \text{ on } B_1 \cup B_2 \cup B_1^+ \cup B_2^+\}, \tag{42}$$

while the coincidence set $A^{st} = \{x \in \Gamma_2 \ : \ u_N = 0\}$. Moreover $B_1 = \{x \in \Gamma_2 : \lambda(x) = -1\}$, $B_2 = \{x \in \Gamma_2 : \lambda(x) = +1\}$, $\tilde{B}_i = \{x \in B_i : u_N(x) = 0\}$, $i = 1, 2$, $B_i^+ = B_i \setminus \tilde{B}_i$, $i = 1, 2$. The derivative of the Lagrangian L with respect to ϕ has the form:

$$\frac{\partial L}{\partial \phi}(\phi,\tilde{\lambda}) = \int_D \rho'(\phi)[a_{ijkl}e_{ij}(u_\epsilon)e_{kl}(p^a+\eta) - f(p^a+\eta)]dx$$
$$+\tilde{\lambda}d'(\phi) + \sum_{i=1}^{3}\frac{1}{\mu_i}d(\phi)d'(\phi), \tag{43}$$

where $\rho'(\phi) = c_1 - c_2 = 1-\epsilon$, $d'(\phi) = [Vol'(\phi), W'(\phi), Per'(\phi)]$ and

$$Vol'(\phi) = 1, \quad W'(\phi) = 2\phi, \tag{44}$$

$$Per'(\phi) = \chi_{\{\partial\Omega = const_0\}}\max\{0, -\nabla\cdot(\frac{\nabla\phi}{\mid\nabla\phi\mid})\} - \chi_{\{\partial\Omega > const_0\}}\nabla\cdot(\frac{\nabla\phi}{\mid\nabla\phi\mid}). \tag{45}$$

Using (39)–(45) we can formulate the necessary optimality condition:

Lemma 3. *If $\hat{\phi} \in U_{ad}^{\phi}$ is an optimal solution to the problem (33)–(37) 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*

$$L(\hat{\phi},\tilde{\lambda}) \leq L(\hat{\phi},\tilde{\lambda}^\star) \leq L(\phi,\tilde{\lambda}^\star) \ \ \forall(\phi,\tilde{\lambda}) \in U_{ad}^{\phi}\times R^3. \tag{46}$$

Proof follows from standard arguments [6,8]. Recall [6,9] condition (46) implies that for all $\phi \in U_{ad}^{\phi}$ and $\tilde{\lambda} \in R^3$

$$\frac{\partial L(\hat{\phi},\tilde{\lambda})}{\partial\phi} \geq 0 \ \text{ and } \ \frac{\partial L(\phi,\tilde{\lambda}^\star)}{\partial\tilde{\lambda}} \leq 0. \tag{47}$$

4 Numerical Experiments

The optimization problem (33)–(37) is discretized using the finite element method [8,9]. The finite difference method is used to approximate interface evolution (gradient flow) equation [2]. The discretized structural optimization problem (33)–(37) is solved numerically. We employ Uzawa type algorithm to solve numerically optimization problem (33)–(37). The algorithm is programmed in Matlab environment. As an example a body occupying 2D domain

$$\Omega = \{(x_1,x_2) \in R^2 : 0 \leq x_1 \leq 8 \ \wedge \ 0 < v(x_1) \leq x_2 \leq 4\}, \tag{48}$$

is considered. The boundary Γ of the domain Ω is divided into three disjoint pieces

$$\begin{aligned}\Gamma_0 &= \{(x_1,x_2) \in R^2 : x_1 = 0,8 \ \wedge \ 0 < v(x_1) \leq x_2 \leq 4\},\\ \Gamma_1 &= \{(x_1,x_2) \in R^2 : 0 \leq x_1 \leq 8 \ \wedge \ x_2 = 4\},\\ \Gamma_2 &= \{(x_1,x_2) \in R^2 : 0 \leq x_1 \leq 8 \ \wedge \ v(x_1) = x_2\}.\end{aligned} \tag{49}$$

The domain Ω and the boundary Γ_2 depend on the function v given as in Fig. 1. The obtained optimal domain is presented in Fig. 2. 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.

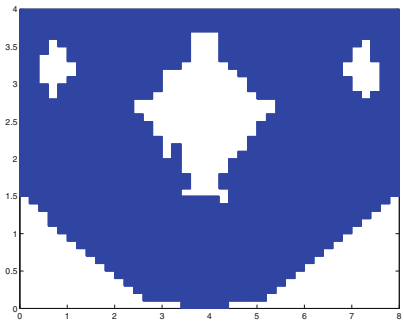


Fig. 2. Optimal domain $\Omega^\star$.

References

1. Allaire, G., Jouve, F., Toader, A.: Structural optimization using sensitivity analysis and a level set method. J. Comput. Phys. **194**, 363–393 (2004)
2. Aubert, G., Kornprobst, P.: Mathematical Problems in Image Processing. Springer, New York (2006)
3. Chambolle, A.: A density result in two-dimensional linearized elasticity and applications. Arch. Ration. Mech. Anal. **167**, 211–233 (2003)
4. Dai, X., Tang, P., Cheng, X., Wu, M.: A variational binary level set method for structural topology optimization. Commun. Comput. Phys. **13**(5), 1292–1308 (2013)
5. Deaton, J.D., Grandhi, R.V.: A survey of structural and multidisciplinary continuum topology optimization: post 2000. Struct. Multidiscipl. Optim. **49**, 1–38 (2014)
6. Delfour, M.C., Zolesio, J.P.: Shapes and Geometries: Analysis, Differential Calculus and Optimization. SIAM Publications, Philadelphia (2001)
7. van Dijk, N.P., Maute, K., Langlaar, M., van Keulen, F.: Level-set methods for structural topology optimization: a review. Struct. Multidiscipl. Optim. **48**, 437–472 (2013)
8. Haslinger, J., Mäkinen, R.: Introduction to Shape Optimization: Theory, Approximation, and Computation. SIAM Publications, Philadelphia (2003)
9. Hlaváček, I., Haslinger, J., Nečas, J., Lovišek, J.: Solving of Variational Inequalities in Mechanics. Springer, New York (1988)
10. Lie, J., Lysaker, M., Tai, X.C.: A binary level set model and some applications to Mumford Shah image segmentation. IEEE Trans. Image Process. **15**(5), 1171–1181 (2006)
11. Myśliński, A.: Level set method for optimization of contact problems. Eng. Anal. Bound. Elem. **32**, 986–994 (2008)
12. 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. Eng. **199**(45–48), 2876–2891 (2010)
13. Osher, S., Fedkiw, R.: Level Set Methods and Dynamic Implicit Surfaces. Springer, New York (2003)

14. Zhu, S., Wu, Q., Liu, C.: Shape and topology optimization for elliptic boundary value problems using a piecewise constant level set method. Appl. Numer. Math. **61**, 752–767 (2011)
15. Zhu, S., Liu, C., Wu, Q.: Binary level set methods for topology and shape optimization of a two-density inhomogeneous drum. Comput. Methods Appl. Mech. Eng. **199**(45–48), 2970–2986 (2010)

Nonlinear Delay Evolution Inclusions on Graphs

Mihai Necula[1], Marius Popescu[2], and Ioan I. Vrabie [1,3(✉)]

[1] Faculty of Mathematics, "Al. I. Cuza" University, 700506 Iaşi, Romania
{necula,ivrabie}@uaic.ro
[2] Department of Mathematics, "Dunărea de Jos" University,
800201 Galaţi, Romania
Marius.Popescu@ugal.ro
[3] Octav Mayer Mathematics Institute, Romanian Academy,
700505 Iaşi, Romania

Abstract. We prove a necessary and a sufficient condition for a time-dependent closed set to be viable with respect to a delay evolution inclusion. An application to a null controllability problem is also included.

Keywords: Delay differential inclusion · m-dissipative operator · Viability · Null controllability problem

1 Introduction

Let X be a real Banach space, $I = [a, b) \subseteq \mathbb{R}$ and let $A : D(A) \subseteq X \rightsquigarrow X$ be the infinitesimal generator of a nonlinear semigroup of nonexpansive mappings $\{S(t) : \overline{D(A)} \to \overline{D(A)};\ t \geq 0\}$. Let $\sigma \geq 0$ and let $C_\sigma = C([-\sigma, 0]; X)$ be endowed with the usual sup-norm $\|\varphi\|_\sigma = \sup\{\|\varphi(t)\|;\ t \in [-\sigma, 0]\}$.

If $u \in C([\tau - \sigma, T], X)$ and $t \in [\tau, T]$, we denote by $u_t \in C_\sigma$ the function defined by $u_t(s) = u(t + s)$ for $s \in [-\sigma, 0]$. It should be noticed that for $\sigma = 0$, i.e. when de delay is absent, C_σ reduces to X. Let $K : I \rightsquigarrow X$ and $F : \mathcal{K} \rightsquigarrow X$ be nonempty-valued multi-functions, where $\mathcal{K} = \{(t, \varphi) \in I \times C_\sigma;\ \varphi(0) \in K(t)\}$.

In this paper prove a necessary and a sufficient condition in order that $\mathcal{K}$ be viable with respect to $A + F$. Let $(\tau, \varphi) \in \mathcal{K}$ and let us consider

$$\begin{cases} u'(t) \in Au(t) + F(t, u_t) \\ u_\tau = \varphi. \end{cases} \tag{1}$$

Definition 1. A function $u \in C([\tau - \sigma, T]; X)$ is said to be a C^0-*solution of* (1) *on* $[\tau, T] \subseteq I$, if $(t, u_t) \in \mathcal{K}$ for $t \in [\tau, T]$, $u(t) = \varphi(t - \tau)$ for $t \in [\tau - \sigma, \tau]$ and there exists $f \in L^1(\tau, T; X)$ with $f(t) \in F(t, u_t)$ a.e. for $t \in [\tau, T]$ and such that u is a C^0-solution of the Cauchy problem

$$\begin{cases} u'(t) \in Au(t) + f(t), \quad t \in [\tau, T] \\ u(\tau) = \varphi(0) \end{cases}$$

in the usual sense. See Cârjă, Necula, Vrabie [2], Definition 1.6.2, p. 17.

Supported by a grant of the Romanian National Authority for Scientific Research, CNCS–UEFISCDI, project number PN-II-ID-PCE-2011-3-0052.

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 210–219, 2014.
DOI: 10.1007/978-3-662-45504-3_20

We say that the function $u : [\tau - \sigma, T) \to X$ is a C^0-*solution of* (1) *on* $[\tau - \sigma, T)$, if u is a C^0-solution on $[\tau - \sigma, \tilde{T}]$ for every $\tilde{T} < T$.

Definition 2. We say that $\mathcal{K}$ is C^0 -*viable* with respect to $A + F$, if for each $(\tau, \varphi) \in \mathcal{K}$, there exists $T > \tau$, such that $[\tau, T] \subseteq I$ and (1) has at least one C^0-solution $u : [\tau - \sigma, T] \to X$. If $T = \sup I$, we say that $\mathcal{K}$ is *globally* C^0*-viable* with respect to $A + F$.

Viability results concerning evolution inclusions without delay, i.e., when $\sigma = 0$, using the concepts of tangent set and quasi-tangent set – introduced and studied by Cârjă, Necula and Vrabie [2–4] and [5] –, were obtained by Necula, Popescu and Vrabie [16,17]. For viability results referring to delay evolution equations and inclusions, we mention the pioneering papers of Pavel and Iacob [18] and Haddad [9]. For related results see Gavioli and Malaguti [8], Lakshmikantham, Leela and Moauro [12], Leela and Moauro [13], Lupulescu and Necula [14]. The semilinear case was very recently considered by Necula and Popescu [15] and the present paper extends to the fully nonlinear case the results there obtained.

The paper is divided into five sections, the second one being concerned with the definitions of the basic concepts used in that follows. In Sect. 3 we state and prove a necessary condition for C^0-viability, while Sect. 4 contains the main result of the paper: a sufficient condition for C^0-viability. In Sect. 5, we include an application to a control problem.

2 Preliminaries

Let $f \in L^1(\tau, T; X)$ and $\xi \in \overline{D(A)}$. We denote by $u(\cdot, \tau, \xi, f) : [\tau, T] \to \overline{D(A)}$ the unique C^0-solution, i.e. integral solution of the Cauchy problem

$$\begin{cases} u'(t) \in Au(t) + f(t), & t \in [\tau, T] \\ u(\tau) = \xi. \end{cases}$$

Clearly, $u(\cdot, \tau, \xi, 0) = S(\cdot - \tau)\xi$, where $\{S(t) : \overline{D(A)} \to \overline{D(A)};\ t \geq 0\}$ is the semigroup of nonexpansive mappings generated by A on $\overline{D(A)}$ by the Crandall and Liggett Exponential Formula. See Crandall and Liggett [7].

We assume familiarity with the basic concepts and results in nonlinear evolution equations, delay equations and inclusions and we refer the reader to Barbu [1], Cârjă, Necula and Vrabie [2], Lakshmikantham and Leela [11], Hale [10] and Vrabie [19] for details.

The metric d on $\mathcal{K}$ is defined by $d((\tau, \varphi), (\theta, \psi)) = \max\{|\tau - \theta|, \|\varphi - \psi\|_\sigma\}$, for all $(\tau, \varphi), (\theta, \psi) \in \mathcal{K}$. Furthermore, whenever we use the term *strongly-weakly u.s.c.* multi-function we mean that the domain of the multi-function in question is equipped with the strong topology, while the range is equipped with the weak topology. The term *u.s.c.* refers to the case in which both domain and range are endowed with the strong, i.e. norm, topology.

Thereafter, $D(\xi, r)$ denotes the closed ball with center ξ and radius r.

Definition 3. The multi-function $F : \mathcal{K} \rightsquigarrow X$ is called *locally bounded* if, for each (τ, φ) in $\mathcal{K}$, there exist $\delta > 0$, $\rho > 0$ and $M > 0$ such that for all (t, ψ) in $([\tau - \delta, \tau + \delta] \times D(\varphi, \rho)) \cap \mathcal{K}$, we have $\|F(t, \psi)\| \leq M$.

Let $(\tau, \varphi) \in \mathcal{K}$, let $\eta \in X$ and let $E \subset X$ be a nonempty, bounded subset, let $h > 0$ and let $\mathcal{F}_E = \left\{ f \in L^1_{\text{loc}}(\mathbb{R}; X);\ f(s) \in E \text{ a.e. for } s \in \mathbb{R} \right\}$. We denote by $u(\tau + h, \tau, \varphi(0), \mathcal{F}_E) = \{u(\tau + h, \tau, \varphi(0), f);\ f \in \mathcal{F}_E\}$.

Definition 4. We say that E is *A-right-quasi-tangent* to $\mathcal{K}$ at (τ, φ) if

$$\liminf_{h \downarrow 0} h^{-1} \mathrm{d}\left(u(\tau + h, \tau, \varphi(0), \mathcal{F}_E), K(\tau + h)\right) = 0.$$

We denote by $\mathcal{QTS}^A_{\mathcal{K}}(\tau, \varphi)$ the set of all A-right-quasi-tangent sets to $\mathcal{K}$ at (τ, φ).

If K is constant, E is right-quasi-tangent to $\mathcal{K}$ at (τ, φ) if and only if it is A-quasi-tangent to K at $\xi = \varphi(0)$ in the sense of Cârjă, Necula, Vrabie [2].

3 Necessary Conditions for Viability

The following lemma was proved in Necula and Popescu [15].

Lemma 1. *Let $f : [\tau, T] \to X$ be a measurable function and $B, C \subset X$ two nonempty sets such that $f(t) \in B + C$ a.e. for $t \in [\tau, T]$. Then, for every $\varepsilon > 0$ there exist $b : [\tau, T] \to B$, $c : [\tau, T] \to C$ and $r : [\tau, T] \to S(0, \varepsilon)$, all measurable, such that $f(t) = b(t) + c(t) + r(t)$* a.e. for $t \in [\tau, T]$.

Theorem 1. *If $F : \mathcal{K} \rightsquigarrow X$ is u.s.c. and $\mathcal{K}$ is C^0-viable with respect to $A + F$ then, for all $(\tau, \varphi) \in \mathcal{K}$, $\lim_{h \downarrow 0} h^{-1} \mathrm{d}\left(u(\tau + h, \tau, \varphi(0), \mathcal{F}_{F(\tau,\varphi)}), K(\tau + h)\right) = 0$.*

Proof. Let $(\tau, \varphi) \in \mathcal{K}$ and $u : [\tau - \sigma, T] \to X$ be a C^0-solution of 1. Hence there exists $f \in L^1(\tau, T; X)$ such that $f(s) \in F(s, u_s)$ a.e. for $s \in [\tau, T]$ and $u(t) = u(t, \tau, \varphi(0), f)$ for all $t \in [\tau, T]$. Let $\varepsilon > 0$ be arbitrary but fixed.

Since F is u.s.c. at (τ, φ) and $\lim_{t \to \tau} u_t = u_\tau = \varphi$ in C_σ, we may find $\delta > 0$ such that $f(s) \in F(s, u_s) \subseteq F(\tau, \varphi) + S(0, \varepsilon)$ a.e. for $s \in [\tau, \tau + \delta]$.

Taking $B = F(\tau, \varphi)$ and $C = S(0, \varepsilon)$, from Lemma 1, we deduce that there exist two integrable functions $g : [\tau, \tau + \delta] \to F(\tau, \varphi)$ and $r : [\tau, \tau + \delta] \to S(0, 2\varepsilon)$ such that $f(s) = g(s) + r(s)$ a.e. for $s \in [\tau, \tau + \delta]$. Since $u(\tau + h) \in K(\tau + h)$, we deduce that, for each $0 < h < \delta$, $\mathrm{d}\left(u(\tau + h, \tau, \varphi(0), \mathcal{F}_{F(\tau,\varphi)}), K(\tau + h)\right)$ $\leq \mathrm{d}\left(u(\tau + h, \tau, \varphi(0), g), u(\tau + h, \tau, \varphi(0), f)\right) \leq \displaystyle\int_\tau^{\tau + h} \|g(s) - f(s)\| ds \leq 2\varepsilon h$. So, $\limsup_{h \downarrow 0} h^{-1} \mathrm{d}\left(u(\tau + h, \tau, \varphi(0), \mathcal{F}_{F(\tau,\varphi)}), K(\tau + h)\right) \leq 2\varepsilon$. The proof is complete.

Theorem 2. *If $F : \mathcal{K} \rightsquigarrow X$ is u.s.c. and $\mathcal{K}$ is C^0-viable with respect to $A + F$ then $F(\tau, \varphi) \in \mathcal{QTS}^A_{\mathcal{K}}(\tau, \varphi)$ for all $(\tau, \varphi) \in \mathcal{K}$.*

4 Sufficient Conditions for Viability

Definition 5. We say that the multi-function $K : I \rightsquigarrow X$ is:

(i) *closed from the left on* I if for any sequence $((t_n, x_n))_{n\geq 1}$ from $I \times X$, with $x_n \in K(t_n)$ and $(t_n)_n$ nondecreasing, $\lim_n t_n = t \in I$ and $\lim_n x_n = x$, we have $x \in K(t)$.
(ii) *locally closed from the left* if for each $(\tau, \xi) \in I \times X$ with $\xi \in K(\tau)$ there exist $T > \tau$ and $\rho > 0$ such that the multi-function $t \rightsquigarrow K(t) \cap D(\xi, \rho)$ is closed from the left on $[\tau, T]$.

Definition 6. An m-dissipative operator $A : D(A) \subseteq X \rightsquigarrow X$ is *of complete continuous type* if for each sequence $(f_n, u_n)_n$ in $L^1(\tau, T; X) \times C([\tau, T]; X)$ with u_n a C^0-solution of the problem $u_n'(t) \in Au_n(t) + f_n(t)$ on $[\tau, T]$ for $n = 1, 2, \ldots$, $\lim_n f_n = f$ weakly in $L^1(\tau, T; X)$ and $\lim_n u_n = u$ strongly in $C([\tau, T]; X)$, it follows that u is a C^0-solution of the problem $u'(t) \in Au(t) + f(t)$ on $[\tau, T]$.

If the dual of X is uniformly convex and A generates a compact semigroup, then A is of complete continuous type. See Vrabie [19, Corollary 2.3.1, p. 49].

Theorem 3. *Let K be locally closed from the left and let $F : \mathcal{K} \rightsquigarrow X$ be nonempty, convex and weakly compact valued. If F is strongly-weakly u.s.c., locally bounded and $A : D(A) \rightsquigarrow X$ is of complete continuous type and generates a compact semigroup, then a sufficient condition in order that $\mathcal{K}$ be C^0-viable with respect to $A + F$ is the tangency condition $F(\tau, \varphi) \in \mathcal{QTS}^A_{\mathcal{K}}(\tau, \varphi)$ for all $(\tau, \varphi) \in \mathcal{K}$. If, in addition, F is u.s.c., then the tangency condition is also necessary in order that $\mathcal{K}$ be C^0-viable with respect to $A + F$.*

The next lemma is inspired from Cârjă and Vrabie [6].

Lemma 2. *Let $K : I \rightsquigarrow X$ be locally closed from the left, $F : \mathcal{K} \rightsquigarrow X$ be locally bounded and let $(\tau, \varphi) \in \mathcal{K}$. Let us assume that the tangency condition is satisfied. Let $\rho > 0$, $T > \tau$ and $M > 0$ be such that:*

(1) *the multi-function $t \rightsquigarrow K(t) \cap D(\varphi(0), \rho)$ is closed from the left on $[\tau, T)$;*
(2) *$\|F(t, \psi)\| \leq M$ for all $t \in [\tau, T]$ and all $\psi \in D_\sigma(\varphi, \rho)$ with $(t, \psi) \in \mathcal{K}$;*
(3) $\sup_{t\in[\tau,T]} \|S(t-\tau)\varphi(0) - \varphi(0)\| + \sup_{|t-s|\leq T-\tau} \|\varphi(t) - \varphi(s)\| + (T-\tau)(M+1) < \rho.$

Then, for each $\varepsilon \in (0, 1)$, there exist a family $\mathcal{P}_T = \{[t_m, s_m); m \in \Gamma\}$ of disjoint intervals, with Γ finite or at most countable, and two functions: $f \in L^1(\tau, T; X)$, and $u \in C([\tau - \sigma, T]; X)$ such that:

(i) $\cup[t_m, s_m) = [\tau, T)$ *and* $s_m - t_m \leq \varepsilon$, *for all* $m \in \Gamma$;
(ii) $u(t_m) \in K(t_m)$, *for all* $m \in \Gamma$ *and* $u(T) \in K(T)$;
(iii) $f(s) \in F(t_m, u_{t_m})$ *a.e. for* $s \in [t_m, s_m)$ *and* $\|f(s)\| \leq M$ *a.e. for* $s \in [\tau, T]$;
(iv) $u(t) = \varphi(t - \tau)$ *for* $t \in [\tau - \sigma, \tau]$ *and* $\|u(t) - u(t, t_m, u(t_m), f)\| \leq (t - t_m)\varepsilon$ *for* $t \in [t_m, T]$ *and* $m \in \Gamma$;
(v) $\|u_t - \varphi\|_\sigma < \rho$ *for all* $t \in [\tau, T]$;
(vi) $\|u(t) - u(t_m)\| \leq \varepsilon$ *for all* $t \in [t_m, s_m)$ *and all* $m \in \Gamma$.

Proof. Let us observe that, if $(i)\sim(iv)$ are satisfied, then (v) is satisfied too, i.e. $\|u(t+s)-\varphi(s)\|<\rho$ for all $t\in[\,\tau,T\,]$ and $s\in[\,-\sigma,0\,]$. Indeed, if $t+s\le\tau$ then

$$\|u(t+s)-\varphi(s)\|=\|\varphi(t+s-\tau)-\varphi(s)\|\le\sup_{|t_1-t_2|\le T-\tau}\|\varphi(t_1)-\varphi(t_2)\|<\rho.$$

If $t+s>\tau$ then $|s|<T-\tau$ and from (3), (iii) and (iv), we get

$$\begin{aligned}\|u(t+s)-\varphi(s)\|&\le\|u(t+s)-u(t+s,\tau,\varphi(0),f)\|\\&+\|u(t+s,\tau,\varphi(0),f)-u(t+s,\tau,\varphi(0),0)\|\\&+\|u(t+s,\tau,\varphi(0),0)-\varphi(0)\|+\|\varphi(0)-\varphi(s)\|\\&\le(t+s-\tau)\varepsilon+\int_\tau^{t+s}\|f(\theta)\|d\theta+\|S(t+s-\tau)\varphi(0)-\varphi(0)\|+\|\varphi(0)-\varphi(s)\|\\&\le(T-\tau)(1+M)+\|S(t+s-\tau)\varphi(0)-\varphi(0)\|+\|\varphi(0)-\varphi(s)\|<\rho.\end{aligned}$$

Let $\varepsilon\in(0,1)$ be arbitrary, but fixed. We will show that there exist $\delta=\delta(\varepsilon)$ in (τ,T) and $\mathcal{P}_\delta$, f, u such that $(i)\sim(vi)$ hold true with δ instead of T.

From the tangency condition, it follows that there exist $h_n\downarrow 0$, $g_n\in\mathcal{F}_{F(\tau,\varphi)}$ and $p_n\in X$, with $\|p_n\|\to 0$ and $u(\tau+h_n,\tau,\varphi(0),g_n)+p_nh_n\in K(\tau+h_n)$ for every $n\in\mathbb{N}$, $n\ge 1$. Let $n_0\in\mathbb{N}$ and $\delta=\tau+h_{n_0}$ be such that $\delta\in(\tau,T)$, $h_{n_0}<\varepsilon$ and $\|p_{n_0}\|<\varepsilon$.

Let $\mathcal{P}_\delta=\{[\,\tau,\delta)\}$, $f(t)=g_{n_0}(t)$ and $u(t)=u(t,\tau,\varphi(0),g_{n_0})+(t-\tau)p_{n_0}$ for $t\in[\,\tau,\delta\,]$. Obviously, $(i)\sim(v)$ are satisfied. Moreover, we may diminish $\delta>\tau$ (increase n_0), if necessary, in order to (vi) be satisfied too.

Let $\mathcal{U}=\{(\mathcal{P}_\delta,f,u);\ \delta\in(\tau,T\,]$ and $(i)\sim(vi)$ are satisfied with δ instead of $T\}$.

As we already have shown, $\mathcal{U}\ne\emptyset$. On $\mathcal{U}$ we define a partial order by:

$$(\mathcal{P}_{\delta_1},f_1,u_1)\preceq(\mathcal{P}_{\delta_2},f_2,u_2),$$

if $\delta_1\le\delta_2$, $\mathcal{P}_{\delta_1}\subseteq\mathcal{P}_{\delta_2}$, $f_1(s)=f_2(s)$ a.e. for $s\in[\,\tau,\delta_1\,]$ and $u_1(s)=u_2(s)$ for all $s\in[\,\tau,\delta_1\,]$. We will prove that each nondecreasing sequence in $\mathcal{U}$ is bounded from above. Let $((\mathcal{P}_{\delta_j},f_j,u_j))_{j\ge1}$ be a nondecreasing sequence in $\mathcal{U}$ and let $\delta=\sup_{j\ge1}\delta_j$. If there exists $j_0\in\mathbb{N}$ such that $\delta_{j_0}=\delta$, then $(\mathcal{P}_{\delta_{j_0}},f_{j_0},u_{j_0})$ is an upper bound for the sequence. So, let us assume that $\delta_j<\delta$, for all $j\ge1$. Obviously, $\delta\in(\tau,T\,]$. We define $\mathcal{P}_\delta=\cup_{j\ge1}\mathcal{P}_{\delta_j}$, $f(t)=f_j(t)$ and $u(t)=u_j(t)$ for all $j\ge1$ and $t\in[\,\tau,\delta_j)$. Clearly, $f\in L^1(\tau,\delta;X)$ and $u\in C([\,\tau,\delta);X)$.

Let us observe that, in view of (iv), we have

$$\begin{aligned}\|u(t)-u(s)\|&\le\|u(t)-u(t,\delta_j,u(\delta_j),f)\|\\&+\|u(t,\delta_j,u(\delta_j),f)-u(s,\delta_j,u(\delta_j),f)\|+\|u(s,\delta_j,u(\delta_j),f)-u(s)\|\\&\le(t-\delta_j)\varepsilon+\|u(t,\delta_j,u(\delta_j),f)-u(s,\delta_j,u(\delta_j),f)\|+(s-\delta_j)\varepsilon\\&\le2(\delta-\delta_j)\varepsilon+\|u(t,\delta_j,u(\delta_j),f)-u(s,\delta_j,u(\delta_j),f)\|\end{aligned}$$

for all $j\ge1$ and all $t,s\in[\delta_j,\delta)$. Since $\lim_j\delta_j=\delta$ and $u(\cdot,\delta_j,u(\delta_j),f)$ is continuous at $t=\delta$, we conclude that u satisfies the Cauchy condition for the

existence of the limit at $t = \delta$. So, u can be extended by continuity to the whole interval $[\tau, \delta]$. By observing that $u(\delta) = \lim\limits_{t \uparrow \delta} u(t) = \lim\limits_{j \to \infty} u(\delta_j) = \lim\limits_{j \to \infty} u_j(\delta_j)$, $u_j(\delta_j) \in D(\varphi(0), \rho) \cap K(\delta_j)$ and the latter is closed from the left, we deduce that $u(\delta) \in D(\varphi(0), \rho) \cap K(\delta)$. The rest of conditions in lemma being obviously satisfied, it follows that $(\mathcal{P}_\delta, f, u)$ is an upper bound for the sequence. Consequently, $(\mathcal{U}, \preceq)$ and $\mathcal{N} : (\mathcal{U}, \preceq) \to R$, defined by $\mathcal{N}(\mathcal{P}_\delta, f, u) = \delta$, for each $(\mathcal{P}_\delta, f, u) \in \mathcal{U}$, satisfy the hypotheses of the Brezis-Browder Ordering Principle – see Cârjă, Necula and Vrabie [2, Theorem 2.1.1, p. 30]. Accordingly, there exists an $\mathcal{N}$-maximal element in $\mathcal{U}$. This means that there exists $(\mathcal{P}_{\delta^*}, f^*, u^*) \in \mathcal{U}$ such that, whenever $(\mathcal{P}_{\delta^*}, f^*, u^*) \preceq (\mathcal{P}_{\overline{\delta}}, \overline{f}, \overline{u})$, we necessarily have $\mathcal{N}(\mathcal{P}_{\delta^*}, f^*, u^*) = \mathcal{N}(\mathcal{P}_{\overline{\delta}}, \overline{f}, \overline{u})$. We will show that $\delta^* = T$. To this aim, let us assume by contradiction that $\delta^* < T$.

Since $(\delta^*, u^*_{\delta^*}) \in \mathcal{K}$, using the tangency condition, we deduce that there exist the sequences $h_n \downarrow 0$, $g_n \in \mathcal{F}_{F(\delta^*, u^*_{\delta^*})}$ and $p_n \in X$, with $\|p_n\| \to 0$, such that $u(\delta^* + h_n, \delta^*, u^*(\delta^*), g_n) + p_n h_n \in K(\delta^* + h_n)$ for all $n \in \mathbb{N}$, $n \geq 1$. Let $n_0 \in \mathbb{N}$ and $\overline{\delta} = \delta^* + h_{n_0}$ with $\overline{\delta} \in (\delta^*, T)$, $h_{n_0} < \varepsilon$ and $\|p_{n_0}\| < \varepsilon$. Let $\mathcal{P}_{\overline{\delta}} = \mathcal{P}_{\delta^*} \cup \{[\delta^*, \overline{\delta}]\}$,

$$\overline{f}(t) = \begin{cases} f^*(t), & t \in [\tau, \delta^*] \\ f_{n_0}(t), & t \in (\delta^*, \overline{\delta}] \end{cases},$$

$$\overline{u}(t) = \begin{cases} u^*(t), & t \in [\tau, \delta^*] \\ u(t, \delta^*, u^*(\delta^*), f_{n_0}) + (t - \delta^*) p_{n_0}, & t \in (\delta^*, \overline{\delta}]. \end{cases}$$

By (v), we have $u^*_{\delta^*} \in S_\sigma(\varphi, \rho)$. So, (2) implies that $\|\overline{f}(s)\| \leq M$ a.e. for $s \in (\tau, \overline{\delta})$.

Clearly $(i) \sim (iii)$ are satisfied. In order to prove (iv) we will consider only the case $t_m \leq \delta^* \leq t$, the other cases being obvious. Using the evolution property, i.e. $u(t, a, \xi, f) = u(t, b, u(b, a, \xi, f), f)$ for $\tau \leq a \leq b \leq t \leq T$, we get

$$\begin{aligned} &\|\overline{u}(t) - u(t, t_m, u^*(t_m), \overline{f})\| \\ &\leq \|u(t, \delta^*, u^*(\delta^*), \overline{f}) - u(t, t_m, u^*(t_m), \overline{f})\| + (t - \delta^*)\varepsilon \\ &= \|u(t, \delta^*, u^*(\delta^*), \overline{f}) - u(t, \delta^*, u(\delta^*, t_m, u^*(t_m), \overline{f}), \overline{f})\| + (t - \delta^*)\varepsilon \\ &\leq \|u^*(\delta^*) - u(\delta^*, t_m, u^*(t_m), \overline{f})\| + (t - \delta^*)\varepsilon \\ &\leq (\delta^* - t_m)\varepsilon + (t - \delta^*)\varepsilon = (t - t_m)\varepsilon, \end{aligned}$$

which proves (iv).

Similarly, we can diminish $\overline{\delta}$ (increase n_0) in order that (vi) be satisfied too.

So, $(\mathcal{P}_{\overline{\delta}}, \overline{f}, \overline{u}) \in \mathcal{U}$, $(\mathcal{P}_{\delta^*}, f^*, u^*) \preceq (\mathcal{P}_{\overline{\delta}}, \overline{f}, \overline{u})$, but $\delta^* < \overline{\delta}$ which contradicts the maximality of $(\mathcal{P}_{\delta^*}, f^*, u^*)$. Hence $\delta^* = T$, and $\mathcal{P}_{\delta^*}$, f^* and u^* satisfy all the conditions $(i) \sim (vi)$. The proof is complete.

Definition 7. Let $\varepsilon > 0$. An element $(\mathcal{P}_T, f, u)$ satisfying $(i) \sim (vi)$ in Lemma 2, is called an *ε-approximate C^0-solution of* (1).

We can proceed now to the proof of Theorem 3.

Proof. The necessity follows from Theorem 2. As long as the proof of the sufficiency is concerned, let $\rho > 0$, $T > \tau$ and $M > 0$ be as in Lemma 2. Let $\varepsilon_n \in (0,1)$, with $\varepsilon_n \downarrow 0$. Let $((\mathcal{P}_T^n, f_n, u_n))_n$ be a sequence of ε_n-approximate C^0-solutions of (1) given by Lemma 2. If $\mathcal{P}_T^n = \{[t_m^n, s_m^n);\ m \in \Gamma_n\}$ with Γ_n finite or at most countable, we denote by $a_n : [\tau, T) \to [\tau, T)$ the step function, defined by $a_n(s) = t_m^n$ for each $s \in [t_m^n, s_m^n)$. Clearly $\lim_n a_n(s) = s$ uniformly for $s \in [\tau, T)$, while from (vi), deduce that $\lim_n \|u_n(t) - u_n(a_n(t))\| = 0$, uniformly for $t \in [\tau, T)$. From (iv), we get

$$\lim_n (u_n(t) - u(t, \tau, \varphi(0), f_n)) = 0 \tag{2}$$

uniformly for $t \in [\tau, T]$. Since $\|f_n(t)\| \le M$ for all $n \in \mathbb{N}$ and a.e. for $t \in [\tau, T]$ and the semigroup generated by A is compact, by Vrabie [19, Theorem 2.3.3, p. 47], we deduce that the set $\{u(\cdot, \tau, \varphi(0), f_n);\ n \ge 1\}$ is relatively compact in $C([\tau, T]; X)$. From this remark and (2), we conclude that $(u_n)_n$ has at least one uniformly convergent subsequence to some function u, subsequence denoted again by $(u_n)_n$.

Since $a_n(t) \uparrow t$, $\lim_n u_n(a_n(t)) = u(t)$, uniformly for $t \in [\tau, T)$ and the mapping $t \to K(t) \cap D(\varphi(0), \rho)$ is closed from the left, we get that $u(t) \in K(t)$ for all $t \in [\tau, T]$. But $\lim_n (u_n)_{a_n(t)} = u_t$ in C_σ, uniformly for $t \in [\tau, T)$. Hence, the set $C = \overline{\{(a_n(t), (u_n)_{a_n(t)}); n \ge 1,\ t \in [\tau, T)\}}$ is compact and $C \subseteq \mathcal{K}$.

At this point, recalling that F is strongly-weakly u.s.c. and has weakly compact values, by Cârjă, Necula and Vrabie [2, Lemma 2.6.1, p. 47], it follows that $B = \overline{\text{conv}}\left(\bigcup_{n\ge1}\bigcup_{t\in[\tau,T)} F(a_n(t), (u_n)_{a_n(t)})\right)$ is weakly compact. We notice that $f_n(s) \in B$ for all $n \ge 1$ and a.e. for $s \in [\tau, T]$. An appeal to Cârjă, Necula and Vrabie [2, Theorem 1.3.8, p. 10] shows that, at least on a subsequence, $\lim_n f_n = f$ weakly in $L^1(\tau, T; X)$. As F is strongly-weakly u.s.c. with closed and convex values while, by Lemma 2, for each $n \ge 1$, we have $f_n(s) \in F(a_n(s), (u_n)_{a_n(s)})$ a.e. for $s \in [\tau, T]$, from Vrabie [19, Theorem 3.1.2, p. 88], we conclude that $f(s) \in F(s, u_s)$ a.e. for $s \in [\tau, T]$.

Finally, by (2) and the fact that A is of complete continuous type, we get $u(t) = u(t, \tau, \varphi(0), f)$ for each $t \in [\tau, T]$ and so, u is a C^0-solution of (1).

Theorem 4. *Let K be closed from the left and let $F : \mathcal{K} \rightsquigarrow X$ be nonempty, convex and weakly compact valued. If there exist $a, b \in C(I)$ such that*

$$\|F(t, \varphi)\| \le a(t) + b(t)\|\varphi(0)\| \quad \text{for all } t \in I \text{ and all } \varphi \in C_\sigma,$$

F is strongly-weakly u.s.c. and $A : D(A) \rightsquigarrow X$ is of complete continuous type and generates a compact semigroup, then a sufficient condition in order that $\mathcal{K}$ be globally C^0-viable with respect to $A + F$ is the tangency condition in Theorem 3. If, in addition, F is u.s.c., then the tangency condition is also necessary in order that $\mathcal{K}$ be mild-viable with respect to $A + F$.

5 A Sufficient Condition for Null Controllability

Let X be a Banach space, $A : D(A) \subseteq X \rightsquigarrow X$ an m-dissipative operator, $g : \mathbb{R}_+ \times C_\sigma \to X$ a given function and $(\tau, \varphi) \in \mathbb{R}_+ \times C_\sigma$ with $\varphi(0) \in \overline{D(A)}$. The problem is how to find a measurable control $c(\cdot)$ taking values in $D(0,1)$ in order to reach the origin in some time T, by C^0-solutions of the state equation

$$\begin{cases} u'(t) \in Au(t) + g(t, u_t) + c(t) \\ u_\tau = \varphi. \end{cases} \tag{3}$$

With $G : \mathbb{R}_+ \times C_\sigma \rightsquigarrow X$, defined by $G(t, v) = av(0) + g(t, v) + D(0,1)$, the above problem reformulates: find $T > 0$ and a C^0-solution of problem

$$\begin{cases} u'(t) \in (A - aI)u(t) + G(t, u_t) \\ u_\tau = \varphi, \quad u(\tau + T) = 0. \end{cases} \tag{4}$$

Theorem 5 and Corollary 1 below are "delay" versions of Cârjă, Necula and Vrabie [3, Theorem 12.1 and Corollary 12.1].

Theorem 5. *Let X be a reflexive Banach space and let $A : D(A) \subseteq X \rightsquigarrow X$ be such that, for some $a \in \mathbb{R}$, $A - aI$ is an m-dissipative operator of complete continuous type and which is the infinitesimal generator of a compact semigroup of contractions, $\{S(t) : \overline{D(A)} \to \overline{D(A)}; \quad t \geq 0\}$. Let $g : \mathbb{R}_+ \times C_\sigma \to X$ be a continuous function such that for some $L > 0$ we have*

$$\|g(t, v)\| \leq L\|v(0)\|, \quad \text{for all } (t, v) \in \mathbb{R}_+ \times C_\sigma. \tag{5}$$

Assume that $0 \in D(A)$ and $0 \in A0$. Then, for each $(\tau, \varphi) \in \mathbb{R}_+ \times C_\sigma$ with $\xi = \varphi(0) \in \overline{D(A)} \setminus \{0\}$, there exists a C^0-solution $u : [\tau, \infty) \to X$ of (4) satisfying

$$\|u(t)\| \leq \|\xi\| - (t - \tau) + (L + a) \int_\tau^t \|u(s)\| ds, \quad \text{for all } t \geq \tau \text{ with } \ u(t) \neq 0. \tag{6}$$

Proof. Let $(\tau, \varphi) \in \mathbb{R}_+ \times C_\sigma$ with $\xi = \varphi(0) \in \overline{D(A)} \setminus \{0\}$. We show that there exist $T \in (0, +\infty)$ and a noncontinuable C^0-solution $(z, u) : [\tau, \tau + T) \to \mathbb{R} \times X$ of the problem

$$\begin{cases} z'(t) = (L + a)\|u(t)\| - 1, & t \in [\tau, \tau + T) \\ u'(t) \in (A - aI)u(t) + G(t, u_t), & t \in [\tau, \tau + T) \\ z_\tau = \|\varphi\| \quad \text{and} \quad u_\tau = \varphi, & \\ \|u(t)\| \leq z(t), & t \in [\tau, \tau + T). \end{cases} \tag{7}$$

On the Banach space $\mathcal{X} = \mathbb{R} \times X$ the operator $\mathcal{A} = (0, A - aI)$ generates a compact semigroup of contractions $\{(1, S(t)); \ (1, S(t)) : \mathbb{R} \times \overline{D(A)} \to \mathcal{X}\}$.

We denote by $\mathcal{C}_\sigma = C([-\sigma, 0]; \mathcal{X}) = C([-\sigma, 0]; \mathbb{R}) \times C([-\sigma, 0]; X)$. Let K be the locally closed set $K = \{(x_1, x_2) \in \mathbb{R}_+ \times (\overline{D(A)} \setminus \{0\}); \ \|x_2\| \leq x_1\}$, with the associate set $\mathcal{K} = \{(t, \psi) \in \mathbb{R} \times \mathcal{C}_\sigma; \ \psi(0) \in K\}$, i.e.

$$\mathcal{K} = \{(t, \psi_1, \psi_2) \in \mathbb{R} \times C([-\sigma, 0]; \mathbb{R}) \times C([-\sigma, 0]; X); \ \|\psi_2(0)\| \leq \psi_1(0)\}$$

and let the multi-function $\mathcal{F} : \mathcal{K} \rightsquigarrow \mathbb{R} \times X$ be defined by

$$\mathcal{F}(t,\psi_1,\psi_2)=((L+a)\|\psi_2(0)\|-1,a\psi_2(0)+g(t,\psi_2)+D(0,1)),\ \text{for}\ (t,\psi_1,\psi_2)\in\mathcal{K}.$$

To show that $\mathcal{F}(\tau,\psi_1,\psi_2)\in\mathcal{QTS}^{\mathcal{A}}_{\mathcal{K}}(\tau,\psi_1,\psi_2)$, for every $(\tau,\psi_1,\psi_2)\in\mathcal{K}$, we shall prove the stronger condition: there exists $(\eta_1,\eta_2)\in\mathcal{F}(\tau,\psi_1,\psi_2)$ such that

$$\liminf_{h\downarrow 0} h^{-1}\mathrm{d}\left(\mathcal{U}(\tau+h,\tau,(\xi_1,\xi_2),(\eta_1,\eta_2)),K\right)=0, \tag{8}$$

where $(\xi_1,\xi_2)=(\psi_1(0),\psi_2(0))$ and $\mathcal{U}(\cdot,\tau,(\xi_1,\xi_2),(\eta_1,\eta_2))$ is the C^0-solution of the corresponding Cauchy problem for the operator $\mathcal{A}$, i.e.

$$\mathcal{U}(t,\tau,(\xi_1,\xi_2),(\eta_1,\eta_2))=(\xi_1+(t-\tau)\eta_1,u(t,\tau,\xi_2,\eta_2))\in\mathcal{X},$$

$u(\cdot,\tau,\xi_2,\eta_2)$ being the corresponding solution for $A-aI$. To this end, it suffices to prove that there exist $(h_n)_n$ in $\mathbb{R}_+$, with $h_n\downarrow 0$, and (θ_n,p_n) in $\mathbb{R}\times X$, with $(\theta_n,p_n)\to(0,0)$, such that, for every $n\in\mathbb{N}$, we have

$$\|u(\tau+h_n,\tau,\xi_2,\eta_2)+h_np_n\|\le\xi_1+h_n\eta_1+h_n\theta_n. \tag{9}$$

Clearly, $\|u\,(\tau+h,\tau,\xi_2,\eta_2)\|\le\|\xi_2\|+\displaystyle\int_\tau^{\tau+h}[\,u(s,\tau,\xi_2,\eta_2),\eta_2\,]_+ds$ for all $h>0$. The normalized semi-inner product, $(x,y)\mapsto[\,x,y\,]_+=\lim\limits_{h\downarrow 0}h^{-1}(\|x+hy\|-\|x\|)$, is u.s.c. Hence, setting $\ell(s):=u(s,\tau,\xi_2,\eta_2)$, we get

$$\liminf_{h\downarrow 0}h^{-1}\int_\tau^{\tau+h}[\ell(s),\eta_2]_+ds\le\limsup_{h\downarrow 0}h^{-1}\int_\tau^{\tau+h}[\,\ell(s),\eta\,]_+ds\le[\,\xi_2,\eta_2\,]_+.$$

Let $\eta_1=(L+a)\|\psi_2(0)\|-1=(L+a)\|\xi_2\|-1$ and $\eta_2=a\xi_2+g(\tau,\psi_2)-\frac{\xi_2}{\|\xi_2\|}$. Clearly, $\eta_2\in a\xi_2+g(\tau,\psi_2)+D(0,1)$ and so, $(\eta_1,\eta_2)\in\mathcal{F}(\tau,\psi_1,\psi_2)$. From (5), we get $[\xi_2,\eta_2]_+=a\|\xi_2\|+[\xi_2,g(\tau,\psi_2)]_+-1\le(L+a)\|\xi_2\|-1=\eta_1$ and hence $\liminf\limits_{h\downarrow 0}h^{-1}\left(\|u(\tau+h,\tau,\xi_2,\eta_2)\|-\|\xi_2\|\right)\le\eta_1$. Keeping in mind that $\|\xi_2\|=\|\psi_2(0)\|\le\psi_1(0)=\xi_1$ since $(\tau,\psi_1,\psi_2)\in\mathcal{K}$, the last inequality proves (9) with $p_n=0$. Thus we get (8). From Theorem 3, $\mathcal{K}$ is C^0-viable with respect to $\mathcal{A}+\mathcal{F}$. As $(\tau,\|\varphi\|,\varphi)\in\mathcal{K}$, thanks to Brezis-Browder Ordering Principle [2, Theorem 2.1.1, p. 30] –, we obtain further that there exist $T\in(0,+\infty\,]$ and a noncontinuable C^0-solution of $(z,u):[\,\tau,\tau+T)\to\mathbb{R}\times X$ of (7) which satisfies $(z(t),u(t))\in K$ for every $t\in[\,\tau,\tau+T)$. This means that (6) is satisfied for every $t\in[\,\tau,\tau+T)$. Since G has sublinear growth, u, as a solution of (4), can be continued to $\mathbb{R}_+$. So, $u(\tau+T)$ exists, even though the solution (z,u) of (7) is defined merely on $[\,\tau,\tau+T)$ if T is finite. In this case, $u(\tau+T)=0$ since otherwise (z,u) can be continued to the right of T which is a contradiction.

Corollary 1. *Under the hypothesis of Theorem 5, the following properties hold.*

(*i*) *If $L+a\le 0$, for any $(\tau,\varphi)\in\mathbb{R}_+\times C_\sigma$ with $\xi=\varphi(0)\in\overline{D(A)}\setminus\{0\}$, there exist a control $c(\cdot)$ and a C^0-solution of (3) that reaches the origin of X in some time $T\le\|\xi\|$ and satisfies $\|u(t)\|\le\|x\|-(t-\tau)$ for all $\tau\le t\le\tau+T$.*

(ii) If $L+a>0$, *for every* $(\tau,\varphi)\in\mathbb{R}_+\times C_\sigma$ *with* $\xi=\varphi(0)\in\overline{D(A)}\setminus\{0\}$ *satisfying* $0<\|\xi\|<1/(L+a)$, *there exist a control* $c(\cdot)$ *and a* C^0*-solution of (3) that reaches the origin of* X *in some time* $T\le(L+a)^{-1}\log\left\{[1-(L+a)\|\xi\|]^{-1}\right\}$ *and* $\|u(t)\|\le e^{(L+a)(t-\tau)}\left[\|\xi\|-(L+a)^{-1}\right]+(L+a)^{-1}$ *for* $t\in[\tau,\tau+T]$.

References

1. Barbu, V.: Nonlinear Differential Equations of Monotone Type in Banach Spaces. Springer Monographs in Mathematics. Springer, New York (2010)
2. Cârjă, O., Necula, M., Vrabie, I.I.: Viability, Invariance and Applications. North-Holland Mathematics Studies, vol. 207. Elsevier, Amsterdam (2007)
3. Cârjă, O., Necula, M., Vrabie, I.I.: Tangent sets, necessary and sufficient conditions for viability for nonlinear evolution inclusions. Set-Valued Anal. **16**, 701–731 (2008)
4. Cârjă, O., Necula, M., Vrabie, I.I.: Necessary and sufficient conditions for viability for semilinear differential inclusions. Trans. Am. Math. Soc. **361**, 343–390 (2009)
5. Cârjă, O., Necula, M., Vrabie, I.I.: Tangent sets, viability for differential inclusions and applications. Nonlinear Anal. **71**, e979–e990 (2009)
6. Cârjă, O., Vrabie, I.I.: Some new viability results for semilinear differential inclusions. NoDEA Nonlinear Differ. Equ. Appl. **4**, 401–424 (1997)
7. Crandall, M.G., Liggett, T.M.: Generation of semi-groups of nonlinear transformations in general Banach spaces. Am. J. Math. **93**, 265–298 (1971)
8. Gavioli, A., Malaguti, L.: Viable mild solutions of differential inclusions with memory in Banach spaces. Port. Math. **57**, 203–217 (2000)
9. Haddad, G.: Monotone trajectories of differential inclusions and functional-differential inclusions with memory. Isr. J. Math. **39**, 83–100 (1981)
10. Hale, J.: Functional differential equations. Applied Mathematical Sciences, vol. 3. Springer, Heidelberg (1971)
11. Lakshmikantham, V., Leela, S.: Nonlinear Differential Equations in Abstract Spaces. International Series in Nonlinear Mathematics, vol. 2. Pergamon Press, New York (1981)
12. Lakshmikantham, V., Leela, S., Moauro, V.: Existence and uniqueness of mild solutions of delay differential equations on a closed subset of a Banach space. Nonlinear Anal. **2**, 311–327 (1978)
13. Leela, S., Moauro, V.: Existence of mild solutions in a closed set for delay differential equations in Banach spaces. Nonlinear Anal. **2**, 47–58 (1978)
14. Lupulescu, V., Necula, M.: A viability result for nonconvex semilinear functional differential inclusions. Discuss. Math. Differ. Incl. Control Optim. **25**, 109–128 (2005)
15. Necula, M. Popescu, M.: A viability result for differential inclusions on graphs, An. Ştiinţ. Univ. Al. I. Cuza Iaşi Secţ. I a Mat., doi:10.2478/aicu-2013-0016
16. Necula, M., Popescu, M., Vrabie, I.I.: Viability for differential inclusions on graphs. Set-Valued Anal. **16**, 961–981 (2008)
17. Necula, M., Popescu, M., Vrabie, I.I.: Evolution equations on locally closed graphs and applications. Nonlinear Anal. **71**, e2205–e2216 (2009)
18. Pavel, N.H., Iacob, F.: Invariant sets for a class of perturbed differential equations of retarded type. Isr. J. Math. **28**, 254–264 (1977)
19. Vrabie, I.I.: Compactness Methods for Nonlinear Evolutions. Pitman Monographs and Surveys in Pure and Applied Mathematics, vol. 75, 2nd edn. Longman, New York (1995)

Graphical Lasso Granger Method with 2-Levels-Thresholding for Recovering Causality Networks

Sergiy Pereverzyev Jr.[1(✉)] and Kateřina Hlaváčková-Schindler[2]

[1] Applied Mathematics Group, Department of Mathematics, University of Innsbruck, Innsbruck, Austria
sergiy.pereverzyev@uibk.ac.at
[2] Chair of Bioinformatics, University of Natural Resources and Life Sciences Vienna, Vienna, Austria
katerina.schindler@gmail.com

Abstract. The recovery of the causality networks with a number of variables is an important problem that arises in various scientific contexts. For detecting the causal relationships in the network with a big number of variables, the so called Graphical Lasso Granger (GLG) method was proposed. It is widely believed that the GLG-method tends to overselect causal relationships. In this paper, we propose a thresholding strategy for the GLG-method, which we call 2-levels-thresholding, and we show that with this strategy the variable overselection of the GLG-method may be overcomed. Moreover, we demonstrate that the GLG-method with the proposed thresholding strategy may become superior to other methods that were proposed for the recovery of the causality networks.

Keywords: Causality network · Gene causality network · Granger causality · Graphical Lasso method · 2-levels-thresholding

1 Introduction

Causality is a relationship between a cause and its effect (its consequence). One can say that *inverse problems* solving, where one would like to discover unobservable features of the cause from the observable features of an effect [4], i.e., searching for the cause of an effect, can in general be seen as a causality problem.

A *causality network* is a directed graph with nodes, which are variables $\{x^j,\ j = 1, \ldots, p\}$, and directed edges, which are the causal influences between the variables. We write $x^i \leftarrow x^j$ if the variable x^j has a causal influence on the variable x^i. Causality networks arise in various scientific contexts.

For example, In Cell Biology one considers causality networks which involve sets of active genes of a cell. An active gene produces a protein. It has been observed that the amount of the protein which is produced by a given gene may depend on, or may be *causally* influenced by, the amount of the proteins which are produced by other genes. In this way, causal relationships between genes

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 220–229, 2014.
DOI: 10.1007/978-3-662-45504-3_21

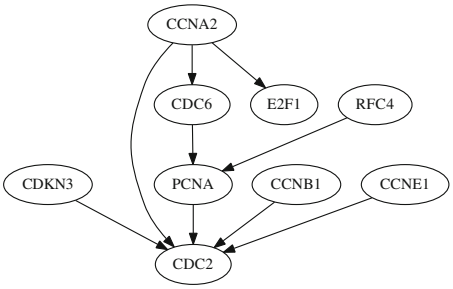


Fig. 1. Causality network of the human cancer cell HeLa genes from the BioGRID database (www.thebiogrid.org).

and the corresponding causality network arise. These causality networks are also called gene regulatory networks. An example of such a network is presented in Fig. 1. This network is achieved from the biological experiments in [9], and it can be found in the BioGRID database. This network has been used in several works [11,14,15] as a test network.

Knowledge of the correct causality networks is important for changing them. In Cell Biology, these networks are used in the research of the causes of genetic diseases. For example, the network in Fig. 1 consists of genes that are active in the human cancer cell HeLa [18]. If one wants to suppress the genes expression in this network, then the primary focus of the suppression therapy should be on the causing genes. For the use of the causality networks in other sciences see, for example, [12].

How can causality network be recovered? In practice, the first information that can be known about the network is the time evolution (time series) of the involved variables $\{x_t^j,\ t = 1,\dots,T\}$. How can this information be used for inferring causal relationships between the variables?

The statistical approach to the derivation of the causal relationships between a variable y and variables $\{z^j,\ j = 1,\dots,p\}$ using the known time evolution of their values $\{y_t, z_t^j,\ t = 1,\dots,T,\ j = 1,\dots,p\}$ consists in considering a model of the relationship between y and $\{z^j,\ j = 1,\dots,p\}$. As a first step, one can consider a linear model of this relationship: $y_t \approx \sum_{j=1}^{p} \beta^j z_t^j, \quad t = 1,\dots,T$. The coefficients $\{\beta^j,\ j = 1,\dots,p\}$ can be specified using the least-squares method. Then, in Statistics [19] by fixing the value of a threshold parameter $\beta_{\rm tr} > 0$, one says that there is a causal relationship $y \leftarrow z^j$ if $|\beta^j| > \beta_{\rm tr}$.

For detecting causal relationships between variables $\{x^j,\ j = 1,\dots,p\}$ the concept of the so called *multivariate Granger causality* has been proposed. This concept originated in the work of Clive Granger [6], who was awarded the Nobel Prize in Economic Sciences in 2003. Based on the intuition that the cause should precede its effect, in Granger causality one says that a variable x^i can be potentially caused by the past versions of the involved variables $\{x^j,\ j = 1,\dots,p\}$.

Then, in the spirit of the statistical approach and using a linear model for the causal relationship, we consider the following approximation problem:

$$x_t^i \approx \sum_{j=1}^{p} \sum_{l=1}^{L} \beta_l^j x_{t-l}^j, \quad t = L+1, \ldots, T, \tag{1}$$

where L is the so called *maximal lag*, which is the maximal number of the considered past versions of the variables. The coefficients $\{\beta_l^j\}$ can be determined by the least-squares method. As in the statistical approach, one can now fix the value of the threshold parameter $\beta_{\text{tr}} > 0$ and say that

$$x^i \leftarrow x^j \quad \text{if} \quad \sum_{l=1}^{L} |\beta_l^j| > \beta_{\text{tr}}. \tag{2}$$

It is well known that for a big number of genes p, as it is pointed out for example in [11], the causality network, which is obtained from the approximation problem (1), is not satisfactory. First of all, it cannot be guaranteed that the solution of the corresponding minimization problem is unique. Another issue is connected with the number of the causality relationships that is obtained from (1). This number is typically very big, while one expects to have a few causality relationships with a given gene. To address this issue, various *variable selection procedures* can be employed. The Lasso [16] is a well known example of such a procedure. In the regularization theory, this approach is known as the l_1-Tikhonov regularization. It has been extensively used for reconstructing the sparse structure of an unknown signal. We refer the interesting reader to [3,5,7,10,13] and the references therein.

The causality concept that is based on the Lasso was proposed in [1] and is named *Graphical Lasso Granger (GLG)* method. However, it is stated in the literature that the Lasso suffers from the variable overselection. And therefore, in the context of the gene causality networks several Lasso modifications were proposed. In [11], the so called *group Lasso* method was considered for recovering gene causality networks using the multivariate Granger causality. The corresponding method can be named Graphical group Lasso Granger (GgrLG) method. And in [15], the *truncating Lasso* method was proposed. The resulting method can be named Graphical truncating Lasso Granger (GtrLG) method.

Nevertheless, it seems that an important tuning possibility of the Lasso, namely an appropriate choice of the *threshold parameter* β_{tr}, has been overlooked in the literature devoted to the recovery of the gene causality networks. In this paper, we are going to show that the GLG-method, which is equipped with an appropriate *thresholding strategy* and an appropriate *regularization parameter choice rule*, may become a superior method in comparison to other methods that were proposed for the recovery of the gene causality networks.

The paper is organized as follows. In Sect. 2, we recall the GLG-method. The quality measures of the graphical methods are presented in Sect. 3. In Sect. 4, we use the network from Fig. 1 to compare the performance of the known graphical methods with the ideal version of the GLG-method, which we call the *optimal*

GLG-estimator. Such a comparison demonstrates the potential of the GLG-approach. In Sect. 5, we propose a thresholding strategy for the GLG-method that allows its automatic realization, which we describe in Sect. 6. Then again we use the network from Fig. 1 to compare the performance of the proposed version of the GLG-method with other graphical methods. It turns out that the proposed method has a superior quality compared to the known methods. The paper is finished with the conclusion and outlook in Sect. 7.

2 Graphical Lasso Granger Method

Let us specify the application of the least-squares method to the approximation problem (1). For this purpose, let us define the vectors $Y^i = (x^i_{L+1}, x^i_{L+2}, \dots, x^i_T)'$, $\boldsymbol{\beta} = (\beta^1_1, \dots, \beta^1_L, \beta^2_1, \dots, \beta^2_L, \dots, \beta^p_1, \dots, \beta^p_L)'$, and the matrix

$$X = \big(\, (x^1_{t-1}, \dots, x^1_{t-L}, x^2_{t-1}, \dots, x^2_{t-L}, \dots, x^p_{t-1}, \dots, x^p_{t-L}); \; t = L+1, \dots, T \, \big).$$

Then, in the least-squares method, one considers the following minimization problem:

$$\|Y^i - X\boldsymbol{\beta}\|^2 \to \min_{\boldsymbol{\beta}}, \tag{3}$$

where $\|\cdot\|$ denotes the l_2-norm.

As it was mentioned in the introduction, the solution of (3) defines unsatisfactory causal relationships and various variable selection procedures should be employed instead. A well-known example of such procedures is the Lasso [16]. In this procedure, one considers the following minimization problem:

$$\|Y^i - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 \to \min_{\boldsymbol{\beta}}. \tag{4}$$

Solution of (4) for each variable $\{x^i, \; i = 1, \dots, p\}$ with the causality rule (2) defines an estimator of the causality network between the variables $\{x^i\}$, and in this way one obtains the Graphical Lasso Granger (GLG) method [1].

3 Quality Measures of the Graphical Methods

A *graphical method* is a method that reconstructs the causality network, which is a directed graph, with the variables $\{x^j\}$. The *quality* of a graphical method can be estimated from its performance on a *known* causality network. The network in Fig. 1 has been used for testing methods' quality in several publications [11, 14,15]. What measures can be used for estimating the quality of a graphical method?

First of all, let us note that a causality network can be characterized by the so called *adjacency matrix* $A = \{A_{i,j} \mid \{i,j\} \subset \{1, \dots, p\}\}$ with the following elements:

$$A_{i,j} = 1 \quad \text{if} \quad x^i \leftarrow x^j; \;\; A_{i,j} = 0 \quad \text{otherwise.}$$

The adjacency matrix A^{true} for the causality network in Fig. 1 is presented in Fig. 2. There, the white squares correspond to $A_{i,j} = 1$, and the black squares—to the zero-elements. The genes are numbered in the following order: CDC2, CDC6, CDKN3, E2F1, PCNA, RFC4, CCNA2, CCNB1, CCNE1.

Now, imagine that there is a *true* adjacency matrix A^{true} of the true causality network, and there is its *estimator* A^{estim}, which is produced by a graphical method. The quality of the estimator A^{estim} can be characterized by the following quality measures: *precision (P)*, *recall (R)*, F_1*-score* (F_1). See, for example, [12] for the detailed definition of these measures.

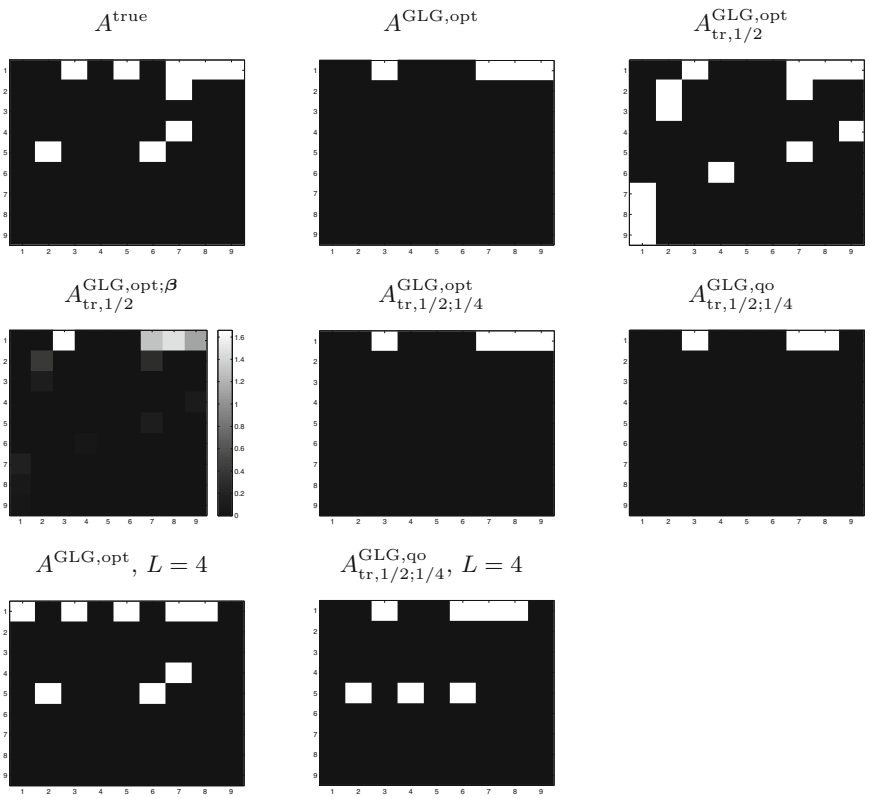


Fig. 2. The adjacency matrix A^{true} for the causality network in Fig. 1 and its various GLG-estimators.

As it was already mentioned, the causality network in Fig. 1 has been used for testing quality of graphical methods. In particular, in [15] one finds the above mentioned quality measures for the following methods: GgrLG, GtrLG and CNET. CNET is a graph search-based algorithm that was introduced in [14]. The data $\{x_i^j\}$ is taken from the third experiment of [18] consisting of

47 time points, and the maximal lag L is taken to be equal to 3. The quality measures from [15] are presented in Table 1.

Table 1. Quality measures of the known graphical methods.

	P	R	F_1
GgrLG	0.24	0.44	0.3
GtrLG	0.3	0.33	0.32
CNET	0.36	0.44	0.4

As it is seen from the table, CNET has the highest F_1-score. However, CNET is the most computationally expensive among the considered methods that does not allow its application to large networks. GgrLG has a good recall but a poor precision, and thus, GtrLG can be considered as a better method among the considered methods.

4 Optimal GLG-estimator

As we have seen in the previous section, the graphical methods, which are based on the Lasso modifications, were tested on the network in Fig. 1 (Table 1). However, the application of the graphical method that is based on the pure Lasso (GLG) to the network in Fig. 1 has not been reported. Moreover, it seems that the possibility of varying the threshold parameter $\beta_{\rm tr}$ in GLG also has not been considered in the literature devoted to the reconstruction of the causality networks.

Assume that the true causality network with the variables $\{x^j\}$ is given by the adjacency matrix $A^{\rm true}$. Assume further that the observation data $\{x^j_t\}$ is given. What is the *best* reconstruction of $A^{\rm true}$ that can be achieved by the GLG-method? The answer to this question is given by, what we call, the *optimal* GLG-estimator. Let us specify its construction.

First of all, let us define the following quality measure, which we call Fs-measure: $Fs = \frac{1}{p^2}\|A^{\rm true} - A^{\rm estim}\|_1$, $0 \le Fs \le 1$. Fs-measure represents the number of *false* elements in the estimator $A^{\rm estim}$ that is scaled with the total number of elements in $A^{\rm estim}$.

Now, let $\boldsymbol{\beta}_i(\lambda)$ denote the solution of the minimization problem (4) in the GLG-method, and $\boldsymbol{\beta}^j_i(\lambda) = (\beta^j_{1,i}, \ldots, \beta^j_{L,i})$. Then, the GLG-estimator $A^{\rm GLG}(\lambda, \beta_{\rm tr})$ of the adjacency matrix $A^{\rm true}$ is defined as follows:

$$A^{\rm GLG}_{i,j}(\lambda, \beta_{\rm tr}) = 1 \quad \text{if} \quad \|\boldsymbol{\beta}^j_i(\lambda)\|_1 > \beta_{\rm tr}; \quad A^{\rm GLG}_{i,j}(\lambda, \beta_{\rm tr}) = 0 \quad \text{otherwise.}$$

The optimal GLG-estimator $A^{\rm GLG,opt}$ of the true adjacency matrix $A^{\rm true}$ is the GLG-estimator $A^{\rm GLG}(\lambda, \beta_{\rm tr})$ with the parameters λ, $\beta_{\rm tr}$ such that the corresponding Fs-measure is minimal.

The optimal GLG-estimator of the adjacency matrix for the causality network in Fig. 1 is presented in Fig. 2. Its quality measures can be found in Table 2. We used the same data $\{x_t^j\}$ as in [11,14,15]. Also, as in [11,15], we take the maximal lag $L = 3$. As one can see, the optimal GLG-estimator reconstructs almost completely the causing genes of the most caused gene in the network. The recall of $A^{\mathrm{GLG,opt}}$ is equal to the highest recall in Table 1, but precision and F_1-score are considerably higher.

Of course, $A^{\mathrm{GLG,opt}}$ is given by the ideal version of the GLG-method, where we essentially use the knowledge of A^{true}. How close can we come to $A^{\mathrm{GLG,opt}}$ without such a knowledge? To answer this question, let us first decide about the choice of the threshold parameter β_{tr}.

5 Thresholding Strategy

The purpose of the threshold parameter β_{tr} is to cancel the causal relationships $x^i \leftarrow x^j$ with *small* $\|\boldsymbol{\beta}_i^j(\lambda)\|_1$. When can we say that $\|\boldsymbol{\beta}_i^j(\lambda)\|_1$ is small? We propose to consider the following *guideindicators* of smallness:

$$\begin{aligned} \beta^i_{\min}(\lambda) &= \min\{\|\boldsymbol{\beta}_i^j(\lambda)\|_1,\ j = 1,\dots,p \mid \|\boldsymbol{\beta}_i^j(\lambda)\|_1 \neq 0\}, \\ \beta^i_{\max}(\lambda) &= \max\{\|\boldsymbol{\beta}_i^j(\lambda)\|_1,\ j = 1,\dots,p\}. \end{aligned} \tag{5}$$

In particular, we propose to consider the threshold parameter of the following form:

$$\beta^i_{\mathrm{tr},\alpha}(\lambda) = \beta^i_{\min}(\lambda) + \alpha(\ \beta^i_{\max}(\lambda) - \beta^i_{\min}(\lambda)). \tag{6}$$

As a default value we take $\alpha = 1/2$.

In the optimal GLG-estimator $A^{\mathrm{GLG,opt}}_{\mathrm{tr},1/2}$ with the threshold parameter $\beta^i_{\mathrm{tr},1/2}$ we choose λ such that the corresponding Fs-measure is minimal. For the causality network in Fig. 1, this estimator is presented in Fig. 2. Its quality measures can be found in Table 2. One observes that although there is some quality decrease in comparison to $A^{\mathrm{GLG,opt}}$, the quality measures are still higher than for the methods in Table 1. However, can this quality be improved?

The choice of the threshold parameter $\beta^i_{\mathrm{tr},1/2}$ rises the following issue. With such a choice we always assign a causal relationship, unless the solution of (4) $\boldsymbol{\beta}_i(\lambda)$ is identically zero. But how strong are these causal relationships compared to each other? The norm $\|\boldsymbol{\beta}_i^j(\lambda)\|_1$ can be seen as a *strongness indicator* of the causal relationship $x^i \leftarrow x^j$.

Let us now construct a matrix $A^{\mathrm{GLG,opt};\boldsymbol{\beta}}_{\mathrm{tr},1/2}$, similarly to the adjacency matrix $A^{\mathrm{GLG,opt}}_{\mathrm{tr},1/2}$, where instead of the element 1 we put the norm $\|\boldsymbol{\beta}_i^j(\lambda)\|_1$, i.e.

$$\begin{aligned} A^{\mathrm{GLG,opt};\boldsymbol{\beta}}_{\mathrm{tr},1/2}(i,j) &= \|\boldsymbol{\beta}_i^j(\lambda^{\mathrm{tr},1/2}_{\mathrm{opt},i})\|_1 \quad \text{if} \quad \|\boldsymbol{\beta}_i^j(\lambda^{\mathrm{tr},1/2}_{\mathrm{opt},i})\|_1 > \beta^i_{\mathrm{tr},1/2}, \\ A^{\mathrm{GLG,opt};\boldsymbol{\beta}}_{\mathrm{tr},1/2}(i,j) &= 0 \quad \text{otherwise.} \end{aligned}$$

This matrix is presented in Fig. 2. One observes that the false causal relationships of the estimator $A^{\mathrm{GLG,opt}}_{\mathrm{tr},1/2}$ are actually weak. This observation suggests to use a second thresholding that is done on the network, or adjacency matrix, level.

We propose to do the thresholding on the network level similarly to the thresholding on the gene level. Namely, let us define the guideindicators of smallness on the network level similarly to (5):

$$A_{\min} = \min\{A^{\mathrm{GLG,opt};\boldsymbol{\beta}}_{\mathrm{tr},1/2}(i,j) \neq 0\},$$
$$A_{\max} = \max\{A^{\mathrm{GLG,opt};\boldsymbol{\beta}}_{\mathrm{tr},1/2}(i,j)\}.$$

And similarly to (6), define the threshold on the network level as follows:

$$A_{\mathrm{tr},\alpha_1} = A_{\min} + \alpha_1(A_{\max} - A_{\min}). \tag{7}$$

We find it suitable to call the described combination of the two thresholdings on the gene and network levels as *2-levels-thresholding*. The adjacency matrix obtained by this thresholding strategy is the following:

$$A^{\mathrm{GLG,opt}}_{\mathrm{tr},1/2;\alpha_1}(i,j) = 1 \quad \text{if} \quad A^{\mathrm{GLG,opt};\boldsymbol{\beta}}_{\mathrm{tr},1/2}(i,j) > A_{\mathrm{tr},\alpha_1},$$
$$A^{\mathrm{GLG,opt}}_{\mathrm{tr},1/2;\alpha_1}(i,j) = 0 \quad \text{otherwise.}$$

It turns out that with $\alpha_1 = 1/4$ in (7) the optimal GLG-estimator can be fully recovered.

6 An Automatic Realization of the GLG-Method

For an automatic realization of the GLG-method, i.e. a realization that does not rely on the knowledge of the true adjacency matrix A^{true}, in addition to a thresholding strategy one needs a choice rule for the regularization parameter λ in (4). For such a choice, we propose to use the so called quasi-optimality criterion [2,8,17]. Some details of the application of this criterion can be found in [12].

The reconstruction obtained by the GLG-method with the 2-levels-thresholding and quasi-optimality criterion $A^{\mathrm{GLG,qo}}_{\mathrm{tr},1/2;1/4}$ is presented in Fig. 2. Its quality measures can be found in Table 2. One observes that there is a little decrease in recall in comparison to the optimal GLG-method; however, this recall is the same as for the GtrLG-method (Table 1). But due to the highest precision, the F_1-score remains to be higher than for the methods in Table 1. Thus, one may say that the proposed realization of the GLG-method outperforms the methods in Table 1.

Nevertheless, one may still wonder, why the proposed realization of the GLG-method captures only the causal relationships of the most caused gene. It appears that the value of the maximal lag L plays an important role in the selection of the causal relationships.

In the modifications of the GLG-method the authors of [11,15] considered $L = 3$. All results presented so far were also obtained with $L = 3$. It turns out that for $L = 4$ the optimal GLG-estimator (see Fig. 2) delivers a much better reconstruction of the causality network. In particular, two more caused genes are recovered.

The proposed automatic realization of the GLG-method with $L = 4$ (Fig. 2) recovers an additional caused gene in comparison to the realization with $L = 3$. Also, all considered quality measures for our automatic realization of the GLG-method with $L = 4$ (Table 2) are considerably higher than for the methods in Table 1. We would like to stress that no use of the knowledge of A^{true} is needed for obtaining $A^{\text{GLG,qo}}_{\text{tr},1/2;1/4}$, and no readjustment of the design parameters α, α_1 is necessary.

Table 2. Quality measures of the various GLG-estimators.

	Fs	P	R	F_1
GLG-opt	6.2 %	1	0.44	0.62
GLG-opt; tr, 1/2	14.8 %	0.38	0.56	0.45
GLG-qo; tr, 1/2; 1/4	7.4 %	1	0.33	0.5
GLG-opt, $L = 4$	3.7 %	0.88	0.78	0.82
GLG-qo; $L = 4$; tr, 1/2; 1/4	7.4 %	0.71	0.56	0.63

7 Conclusion and Outlook

The proposed realization of the Graphical Lasso Granger method with 2-levels-thresholding and quasi-optimality criterion for the choice of the regularization parameter shows a considerable improvement of the reconstruction quality in comparison to other graphical methods. So, the proposed realization is a very promising method for recovering causality networks. Further tests and developments of the proposed realization are worthwhile. In particular, applications to larger causality networks are of interest.

As an open problem for the future, one could consider a study of the choice of the maximal lag and its possible variation with respect to the caused and causing genes.

References

1. Arnold, A., Liu, Y., Abe, N.: Temporal causal modeling with graphical Granger methods. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 66–75. ACM, New York (2007)
2. Bauer, F., Reiß, M.: Regularization independent of the noise level: an analysis of quasi-optimality. Inverse Probl. **24**(5), 16 (2008)
3. Daubechies, I., Defrise, M., De Mol, C.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Commun. Pure Appl. Math. **57**(11), 1413–1457 (2004)
4. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of inverse problems. Kluwer Academic Publishers, Dordrecht (1996)

5. Fornasier, M. (ed.): Theoretical Foundations and Numerical Methods for Sparse Recovery. de Gruyter, Berlin (2010)
6. Granger, C.: Investigating causal relations by econometric models and crossspectral methods. Econometrica **37**, 424–438 (1969)
7. Grasmair, M., Haltmeier, M., Scherzer, O.: Sparse regularization with l^q penalty term. Inverse Probl. **24**(5), 13 (2008)
8. Kindermann, S., Neubauer, A.: On the convergence of the quasioptimality criterion for (iterated) Tikhonov regularization. Inverse Probl. Imaging **2**(2), 291–299 (2008)
9. Li, X., et al.: Discovery of time-delayed gene regulatory networks based on temporal gene expression profiling. BMC Bioinform. **7**, 26 (2006)
10. Lorenz, D.A., Maass, P., Pham, Q.M.: Gradient descent for Tikhonov functionals with sparsity constraints: theory and numerical comparison of step size rules. Electron. Trans. Numer. Anal. **39**, 437–463 (2012)
11. Lozano, A.C., Abe, N., Liu, Y., Rosset, S.: Grouped graphical Granger modeling for gene expression regulatory networks discovery. Bioinformatics **25**, 110–118 (2009)
12. Pereverzyev, S., Jr., Hlaváčková-Schindler, K.: Graphical Lasso Granger method with 2-levels-thresholding for recovering causality networks. Technical report, Applied Mathematics Group, Department of Mathematics, University of Innsbruck (2013)
13. Ramlau, R., Teschke, G.: A Tikhonov-based projection iteration for nonlinear ill-posed problems with sparsity constraints. Numer. Math. **104**(2), 177–203 (2006)
14. Sambo, F., Camillo, B.D., Toffolo, G.: CNET: an algorithm for reverse engineering of causal gene networks. In: NETTAB2008, Varenna, Italy (2008)
15. Shojaie, A., Michailidis, G.: Discovering graphical Granger causality using the truncating lasso penalty. Bioinformatics **26**, i517–i523 (2010)
16. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B **58**, 267–288 (1996)
17. Tikhonov, A.N., Glasko, V.B.: Use of the regularization method in non-linear problems. USSR Comp. Math. Math. Phys. **5**, 93–107 (1965)
18. Whitfield, M.L., et al.: Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol. Biol. Cell **13**, 1977–2000 (2002)
19. Wikipedia. Causality – Wikipedia. The Free Encyclopedia (2013). Accessed 11 Oct 2013

Right-Hand Side Dependent Bounds for GMRES Applied to Ill-Posed Problems

Jennifer Pestana(✉)

Mathematical Institute, University of Oxford, Andrew Wiles Building, Radcliffe Observatory Quarter, Woodstock Road, Oxford OX2 6GG, UK
pestana@maths.ox.ac.uk

Abstract. In this paper we apply simple GMRES bounds to the nearly singular systems that arise in ill-posed problems. Our bounds depend on the eigenvalues of the coefficient matrix, the right-hand side vector and the nonnormality of the system. The bounds show that GMRES residuals initially decrease, as residual components associated with large eigenvalues are reduced, after which semi-convergence can be expected because of the effects of small eigenvalues.

Keywords: GMRES · Convergence · Ill-posed problem

1 Introduction

The solution of an ill-posed problem often requires the solution of a large, sparse linear system $A\boldsymbol{x} = \boldsymbol{b}$ where $A \in \mathbb{C}^{n\times n}$ is non-Hermitian and nearly singular, $\boldsymbol{b} \in \mathbb{C}^n$ and $\boldsymbol{b} \in \text{range}(A)$ [1]. We assume throughout that A is diagonalizable since, although possible, analysis using the Jordan canonical form is more complicated. The near-singularity of A is reflected in a number of small eigenvalues.

In many cases $\boldsymbol{b}$ is unknown and we instead possess a noisy vector $\boldsymbol{b}_\delta$, where $\|\boldsymbol{b} - \boldsymbol{b}_\delta\|_2 = \delta$. This is problematic since the ill-conditioning of A means that $A^{-1}\boldsymbol{b}_\delta$ may be a poor approximation of $\boldsymbol{x}$. Consequently, it is necessary to regularize, i.e., to solve

$$A_\delta \boldsymbol{x}_\delta = \boldsymbol{b}_\delta. \tag{1}$$

The Generalized Minimal Residual method [2] (GMRES) is an iterative method for solving (1) that, given an initial guess $\boldsymbol{x}_0$ which we assume for simplicity is the zero vector, selects at the kth step the iterate $\boldsymbol{x}_k$ for which the residual $\boldsymbol{r}_k = \boldsymbol{b}_\delta - A_\delta \boldsymbol{x}_k$ satisfies

$$\|\boldsymbol{r}_k\|_2 = \min_{\substack{q\in\Pi_k \\ q(0)=1}} \|q(A_\delta)\boldsymbol{b}_\delta\|_2, \tag{2}$$

where Π_k is the set of polynomials of degree at most k. When GMRES is used to solve (1) it can sometimes give good approximations to $\boldsymbol{x}_\delta$ as long as the method

This publication is based on work supported by Award No. KUK-C1-013-04, made by King Abdullah University of Science and Technology (KAUST).

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 230–236, 2014.
DOI: 10.1007/978-3-662-45504-3_22

is terminated after the correct number of iterations, i.e., GMRES itself can have a regularizing effect [3,4]. Alternatively, regularization may be achieved by preconditioning [4–6]. In either case it is important to understand the behaviour of GMRES applied to nearly singular systems. Eldén and Simoncini [4] used the Schur decomposition to show that when the right-hand side has leading components in the direction of eigenvectors associated with large eigenvalues, the initial convergence is related to a reduction in the sizes of these components. Here we provide a complementary analysis involving the eigenvalue-eigenvector decomposition and the simple bounds in Titley-Peloquin, Pestana and Wathen [7]. Similarly to Eldén and Simoncini we find that the first phase of convergence is related to large eigenvalues. We additionally observe that the stagnation typically observed in the second phase, known as semi-convergence, is attributable to the remaining small eigenvalues.

2 Structure of Nearly Singular Systems

Let A_δ have diagonalization $A_\delta = Z\Lambda Z^{-1}$, $\Lambda = \mathrm{diag}(\lambda_i)$ and $Z \in \mathbb{C}^{n\times n}$, where without loss of generality $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$. We wish to separate the spectrum of A_δ into p large eigenvalues and the remaining small eigenvalues. The matrix A_δ may have two distinct sets of eigenvalues, for example, when a preconditioner is applied. In other cases, however, there is no obvious separation. In this situation we find that a division on the order of δ is a reasonable choice.

Given these two sets of eigenvalues we partition A_δ as

$$A_\delta = \begin{bmatrix} Z_1 & Z_2 \end{bmatrix} \begin{bmatrix} \Lambda_1 & \\ & \Lambda_2 \end{bmatrix} \begin{bmatrix} Y_1^* \\ Y_2^* \end{bmatrix},$$

where $\Lambda_1 \in \mathbb{C}^{p\times p}$, $\Lambda_2 \in \mathbb{C}^{(n-p)\times(n-p)}$, $Z_1, Y_1 \in \mathbb{C}^{n\times p}$ and $Z_2, Y_2 \in \mathbb{C}^{n\times(n-p)}$. We assume that $\|Y_2^*\boldsymbol{b}\|_2 = \epsilon$ is small, i.e., that the true right-hand side vector $\boldsymbol{b}$ is mainly associated with the low-frequency components of A_δ; otherwise the ill-posed problem is intractable.

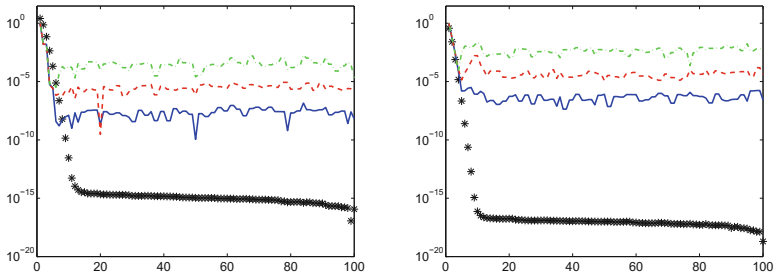
Integral to our bounds are the co-ordinates of $\boldsymbol{b}_\delta$ in the eigenvector basis

$$\boldsymbol{w} = Z^{-1}\boldsymbol{b}_\delta / \|\boldsymbol{b}_\delta\|_2 = \begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{bmatrix} = \frac{1}{\|\boldsymbol{b}_\delta\|_2} \begin{bmatrix} Y_1^*\boldsymbol{b}_\delta \\ Y_2^*\boldsymbol{b}_\delta \end{bmatrix} \tag{3}$$

and, in particular, $\boldsymbol{w}_2 = (Y_2^*\boldsymbol{b} + Y_2^*(\boldsymbol{b}_\delta - \boldsymbol{b}))/\|\boldsymbol{b}_\delta\|_2$, the norm of which is bounded by

$$\|\boldsymbol{w}_2\|_2 \leq (\epsilon + \delta\|Y_2\|_2)/\|\boldsymbol{b}_\delta\|_2. \tag{4}$$

To give some idea of typical spectra, and to show the difference between the components of $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$, we compute these quantities for the baart and wing test problems from the Matlab toolbox Regularization Tools [8,9]. The problems are described in more detail in Sect. 4. We add Gaussian noise to the true right-hand side vectors with $\delta = 10^{-7}$, 10^{-5} and 10^{-3}. For baart, $\|\boldsymbol{b}_\delta\|_2 \approx 2.9$, $\|Y_2\|_2 = 64$ and $\epsilon = 10\delta$ when $p = 5$. Thus, (4) gives $\|\boldsymbol{w}_2\|_2 \leq 26\delta$



(a) baart (b) wing

Fig. 1. Magnitudes of eigenvalues (∗) of A_δ and of corresponding components of $\boldsymbol{w}$ for $\delta = 10^{-7}$ (solid line) $\delta = 10^{-5}$ (dashed line) and $\delta = 10^{-3}$ (dot-dashed line).

for baart. For wing, $\|\boldsymbol{b}_\delta\|_2 \approx 0.15$ and $\|Y_2\|_2 = 158$ with $p = 3$. We find that when δ is 10^{-7}, 10^{-5} and 10^{-3}, ϵ is 1×10^{-5}, 3.6×10^{-4} and 9×10^{-3}, so that (4) is 2×10^{-4}, 0.01 and 1.

Figure 1 shows that for both problems, as expected, the eigenvalues decay and there are a number of very small eigenvalues present. Associated with large eigenvalues are relatively large components of $\boldsymbol{w}$ in magnitude. Once the eigenvalues decrease to around the level of the noise, the components of $\boldsymbol{w}$ stay constant in magnitude at a level that depends on $\|\boldsymbol{b}_\delta\|_2$, the amount of noise and the conditioning of the eigenvectors associated with small eigenvalues. This level is, consequently, higher for wing than for baart. The structure of these two systems is typical of ill-posed linear systems and is exploited in the next section to analyse the convergence of GMRES.

3 GMRES Bounds

Our interest is in explaining the behaviour of GMRES applied to (1). To this end, we apply the bounds in Sect. 2 in Titley-Peloquin *et al.* [7], the first of which is cast in terms of a weighted least squares problem.

Theorem 1. *Suppose that A_δ has diagonalization $A_\delta = Z\Lambda Z^{-1}$, $\Lambda = \mathrm{diag}(\lambda_i)$, and let $\boldsymbol{w}_1 = W_1\boldsymbol{e}$ and $\boldsymbol{w}_2 = W_2\boldsymbol{e}$, where $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are as in (3), $W = \mathrm{diag}(w_i)$ and $\boldsymbol{e} = [1, \ldots, 1]^T$. Then the GMRES residuals satisfy*

$$\frac{\|\boldsymbol{r}_k\|_2}{\|\boldsymbol{b}\|_2} \leq \|Z\|_2 \min_{\substack{q \in \Pi_k \\ q(0)=1}} \left\| \begin{bmatrix} W_1 q(\Lambda_1) & \\ & W_2 q(\Lambda_2) \end{bmatrix} \boldsymbol{e} \right\|_2 . \tag{5}$$

For our ill-posed problem, the weights in W_1 are larger in magnitude than those in W_2 and the eigenvalues in Λ_1 are all larger in magnitude than the eigenvalues in Λ_2. Thus, GMRES will initially choose polynomials that primarily reduce the size of $W_1 q(\Lambda_1)$ to the size of $W_2 q(\Lambda_2)$. In particular, when $\|\boldsymbol{w}_1\|_2 \gg$

$\|\boldsymbol{w}_2\|_2$ we would expect that for the first p steps GMRES would mainly work on reducing the components of the residual associated with Λ_1 and Z_1.

When $\|W_1 q(\Lambda_1)\|_2$ is on the order of $\|W_2(\Lambda_2)\|_2$ it is common for convergence to stagnate, after which residuals may increase in norm; this is known as semi-convergence. The following theorem can help to explain why semi-convergence occurs by explicitly separating the effects of large and small eigenvalues [7].

Theorem 2. *Let A_δ have diagonalization $A_\delta = Z\Lambda Z^{-1}$. For any subset of indices $\mathcal{J}$ with $|\mathcal{J}| = p$, GMRES residuals with $k > p$ satisfy*

$$\frac{\|\boldsymbol{r}_k\|_2}{\|\boldsymbol{b}_\delta\|_2} \leq \|Z\|_2 \min_{\substack{q \in \Pi_{k-p} \\ q(0)=1}} \left(\sum_{\substack{i=1 \\ i \notin \mathcal{J}}}^{n} |\tilde{w}_i|^2 |q(\lambda_i)|^2 \right)^{1/2}, \tag{6}$$

where

$$\tilde{w}_i = w_i \prod_{j \in \mathcal{J}} \left(1 - \frac{\lambda_i}{\lambda_j}\right).$$

To examine the semi-convergence phase, we choose $\mathcal{J} = [1, p]$. Then for any $i \in [p+1, n]$, we have that $|\lambda_i| \leq |\lambda_j|$ and $|\tilde{w}_i| \leq \alpha^p |w_i|$, where $\alpha \leq 2$ and α is around 1 or smaller when, say, there is a decent gap between the large and small eigenvalues or when all eigenvalues have the same sign. Thus, for any $k > p$

$$\frac{\|\boldsymbol{r}_k\|_2}{\|\boldsymbol{b}_\delta\|_2} \leq \alpha^p \|Z\|_2 \|\boldsymbol{w}_2\|_2 \min_{\substack{q \in \Pi_{k-p} \\ q(0)=1}} \|q(\Lambda_2)\|_2.$$

Now, let us consider $\|q(\Lambda_2)\|_2$. Since $|\lambda_i| \ll 1$, $i = p+1, \ldots, n$ and $q(0) = 1$ it will be difficult to reduce $\|q(\Lambda_2)\|_2$ significantly below 1. Consequently, we expect the residuals to stagnate at a level bounded by

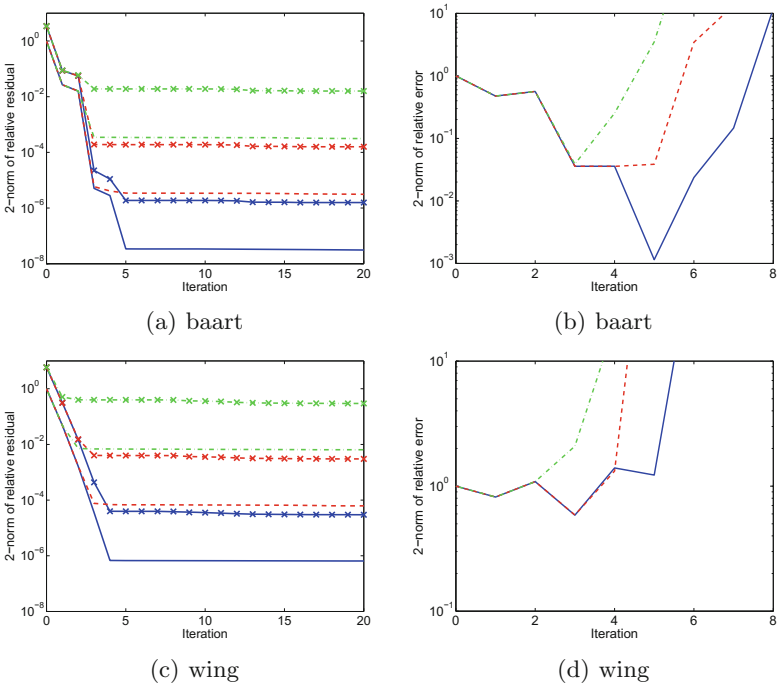
$$\frac{\|\boldsymbol{r}_k\|_2}{\|\boldsymbol{b}_\delta\|_2} \leq \alpha^p \|Z\|_2 \|\boldsymbol{w}_2\|_2. \tag{7}$$

This, in conjunction with (4), indicates that the level of semi-convergence depends on the sizes of the large and small eigenvalues, the noise level δ, the norm of $\boldsymbol{b}_\delta$ and the conditioning of the eigenvectors associated with small eigenvalues.

4 Numerical Results

We now compare the bounds (5) and (7) to the GMRES residuals for the baart and wing problems mentioned above, both of which are discretizations of Fredholm integral equations of the first kind. The integral equation for baart is

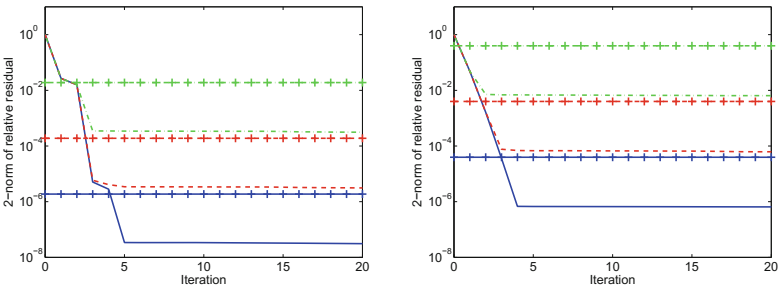
$$\int_0^\pi e^{s\cos(t)} f(t) dt = 2\frac{\sinh(s)}{s}, \quad 0 \leq s \leq \frac{\pi}{2},$$



(a) baart (b) baart

(c) wing (d) wing

Fig. 2. Plots of the relative GMRES residuals and (5) (×) (left) and relative errors (right) for $\delta = 10^{-7}$ (solid line) $\delta = 10^{-5}$ (dashed line) and $\delta = 10^{-3}$ (dot-dashed line).



(a) baart (b) wing

Fig. 3. Plots of the relative GMRES residuals and (7) (+) for $\delta = 10^{-7}$ (solid line) $\delta = 10^{-5}$ (dashed line) and $\delta = 10^{-3}$ (dot-dashed line).

which has the continuous solution $f(t) = \sin(t)$. For the wing problem we solve

$$\int_0^1 te^{-st^2} f(t)dt = \frac{e^{-st_1^2} - e^{-st_2^2}}{2s}, \quad 0 \leq s \leq 1,$$

with $t_1 = 1/3$ and $t_2 = 2/3$. The discontinuous solution is

$$f(t) = \begin{cases} 1 & t_1 < t < t_2, \\ 0 & \text{elsewhere} . \end{cases}$$

Figure 2 shows the relative GMRES residuals and the relative errors. For both baart and wing the relative residuals decrease before stagnating at a level related to the noise level δ. Note that the staircase-like convergence behaviour for baart is particular to this problem. It appears to be related to the harmonic Ritz values, which at the kth step of GMRES are the eigenvalues of a certain $k \times k$ matrix, and which define the GMRES polynomial q in (2) [10]. For fast convergence it is desirable that these harmonic Ritz values are good approximations of eigenvalues of A. For baart, however, at the second and fourth steps there is a harmonic Ritz value that lies between two consecutive eigenvalues of A; these are precisely the steps at which there is little reduction in the relative residual norm.

Unlike the relative residuals, for both problems the norm of the error initially decreases but then starts to increase. This increase occurs during the semi-convergence phase for baart but for the wing problem the errors increase before semi-convergence and exhibit a sawtooth-like behaviour. This highlights the importance of applying a sensible stopping criterion and the potential unsuitability of standard (unpreconditioned) GMRES for some ill-posed problems. Interestingly, (5) seems to provide a better indication of when the iterations should be stopped than the onset of semi-convergence for the wing problem for noisy right-hand side vectors, although we have not investigated this further.

It is clear from Fig. 2 that the bound (5) is very descriptive during the first phase of convergence. Although the bound is not quantitatively descriptive in the second phase of convergence, it accurately predicts the onset of semi-convergence. The approximation (7) is an upper bound on the relative residuals during the semi-convergence phase for both problems (see Fig. 3). Note that for both problems $\alpha \approx 1$. Since (6) is an upper bound on (5), we cannot expect (7) to be quantitatively accurate. Nevertheless, it provides an analysis of semi-convergence and the factors that can affect the level at which residual norms stagnate.

5 Conclusions

In this paper we have applied simple bounds on GMRES convergence to the nearly singular systems that arise from ill-posed problems. We have shown that GMRES initially reduces the residual components associated with large eigenvalues. Once these components are commensurate with those associated with small eigenvalues semi-convergence sets in, with the level at which residuals stagnate determined by the sizes of small eigenvalues, the noise in the right-hand side vector, the size of $\boldsymbol{b}$ and the eigenvectors associated with small eigenvalues.

References

1. Hansen, P.C.: Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion. SIAM, Philadelphia (1998)
2. Saad, Y., Schultz, M.H.: GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. **7**, 856–869 (1986)
3. Calvetti, D., Lewis, B., Reichel, L.: On the regularizing properties of the GMRES method. Numer. Math. **91**, 605–625 (2002)
4. Eldén, L., Simoncini, V.: Solving ill-posed linear systems with GMRES and a singular preconditioner. SIAM J. Matrix Anal. Appl. **33**, 1369–1394 (2012)
5. Hanke, M., Nagy, J.: Inverse Toeplitz preconditioners for ill-posed problems. Linear Algebra Appl. **284**, 137–156 (1998)
6. Hansen, P.C., Jensen, T.K.: Smoothing-norm preconditioning for regularizing minimum-residual methods. SIAM J. Matrix Anal. Appl. **29**, 1–14 (2006)
7. Titley-Peloquin, D., Pestana, J., Wathen, A.J.: GMRES convergence bounds that depend on the right-hand-side vector. IMA J. Numer. Anal. **34**, 462–479 (2014)
8. Hansen, P.C.: Regularization tools version 4.0 for Matlab 7.3. Numer. Algorithms **46**, 189–194 (2007)
9. Hansen, P.C.: Regularization Tools Version 4.1 (for Matlab Version 7.3)
10. Goossens, S., Roose, D.: Ritz and harmonic Ritz values and the convergence of FOM and GMRES. Numer. Linear Algebra Appl. **6**, 281–293 (1997)

PDE-Driven Shape Optimization: Numerical Investigation of Different Descent Directions and Projections Using Penalization and Regularization

Peter Philip(✉)

Department of Mathematics, Ludwig-Maximilians University (LMU) Munich,
Theresienstrasse 39, 80333 Munich, Germany
philip@math.lmu.de
http://www.math.lmu.de/philip

Abstract. We consider shape optimization problems with elliptic partial differential state equations.Using regularization and penalization, unknown shapes are encoded via shape functions, turning the shape optimization into optimal control problems for the unknown functions. The method is designed to allow topological changes in a natural way. Based on convergence and differentiability results, numerical algorithms are formulated, using different descent directions and projections. The algorithms are assessed in a series of numerical experiments, applied to an elliptic PDE arising from an oil industry application with two unknown shapes, one giving the region where the PDE is solved, and the other determining the PDE's coefficients.

Keywords: Shape optimization · Optimal control · Fixed domain method · Elliptic partial differential equation · Numerical simulation

1 Introduction

We study an elliptic shape optimization problem motivated by the oil industry application studied in [12], where one aims at monitoring the interior of a pipeline. The cross section through the pipeline is modeled by a set $D \subseteq \mathbb{R}^2$, consisting of a liquid region Ω (such that $D \setminus \Omega$ represents air) that is part oil (region $O \subseteq \Omega$) and part water ($\Omega \setminus O$). The shape optimization needs to reconstruct the parts O and Ω from given measurements y_d of a quantity (e.g. voltage) taken in a region E adjacent to the boundary of D. Thus, we are led to the following problem, previously formulated in [5]:

$$\min_{\Omega, O} \frac{1}{2} \int_E |y - y_d|^2 \, \mathrm{d}x + \frac{1}{2} \int_E |\nabla y - \nabla y_d|^2 \, \mathrm{d}x \,, \tag{1a}$$

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 237–246, 2014.
DOI: 10.1007/978-3-662-45504-3_23

subject to

$$\int_\Omega \big[a_1\,\chi_O + a_2(1-\chi_O)\big]\nabla y\cdot\nabla v\,\mathrm{d}x \;+\; \int_\Omega \big[b_1\,\chi_O + b_2(1-\chi_O)\big]\,y\,v\,\mathrm{d}x$$
$$= \int_\Omega f\,v\,\mathrm{d}x\,,\quad \forall\, v\in H_0^1(\Omega), \tag{1b}$$

$$y-\xi\in H_0^1(\Omega), \tag{1c}$$

where the sets $E\subseteq O\subseteq \Omega\subseteq D\subseteq\mathbb{R}^2$ all are bounded and open, D also connected; χ_O denotes the characteristic function of O; $a_1,a_2,b_1,b_2>0$, $f\in L^2(D)$, $y_d,\xi\in H^1(D)$ all given. In particular, depending on ξ, (1c) can mean homogeneous or nonhomogeneous Dirichlet conditions.

The employed method (previously used, e.g., in [4,5,8]) is based on the introduction of shape functions g and p, defined on $\overline{D}$ and encoding the unknown sets Ω and O, respectively, and on a technique for the approximation and regularization of characteristic functions. Assuming $g,p:\overline{D}\to\mathbb{R}$ to be continuous, the corresponding sets are obtained via

$$\Omega_g=\mathrm{int}\{x\in D:\,g(x)\geq 0\},\quad O_p=\mathrm{int}\{x\in D:\,p(x)\geq 0\} \tag{2}$$

(Ω_g and O_p then are open Caratheodory sets, not necessarily connected). Enforcing the constraints $E\subseteq O\subseteq\Omega$ translates into the set of admissible pairs

$$U_{\mathrm{ad}}:=\big\{(g,p)\in C(\overline{D})\times C(\overline{D}):\,g\geq p \text{ on } D \text{ and } p\geq 0 \text{ on } E\big\}. \tag{3}$$

Let $H:\mathbb{R}\to\mathbb{R}$ denote the Heaviside function. Then $H(g),H(p):\overline{D}\to\mathbb{R}$ are the characteristic functions of $\overline{\Omega}_g$ and $\overline{O}_p$, respectively. We use the differentiable regularization of the Heaviside function given by

$$H_\varepsilon(r):=\begin{cases}1 & \text{for } r\geq 0,\\ \frac{\varepsilon(r+\varepsilon)^2-2r(r+\varepsilon)^2}{\varepsilon^3} & \text{for } -\varepsilon<r<0,\\ 0 & \text{for } r\leq-\varepsilon,\end{cases} \tag{4}$$

and obtain the following regularized fixed domain approximation of (1) (with the abovementioned constraints $E\subseteq O\subseteq\Omega$):

$$\min_{(g,p)\in U_{\mathrm{ad}}}\ \frac{1}{2}\int_E |y_\varepsilon-y_d|^2\,\mathrm{d}x\;+\frac{1}{2}\int_E|\nabla y_\varepsilon-\nabla y_d|^2\,\mathrm{d}x\,, \tag{5a}$$

$$\int_D\Big[\big[a_1\,H_\varepsilon(p)+a_2(1-H_\varepsilon(p))\big]\nabla y_\varepsilon\cdot\nabla v+\big[b_1\,H_\varepsilon(p)+b_2(1-H_\varepsilon(p))\big]\,y_\varepsilon v\Big]\,\mathrm{d}x$$
$$+\frac{1}{\varepsilon}\int_D\big(1-H_\varepsilon(g)\big)\,y_\varepsilon v\,\mathrm{d}x\;=\int_D f v\,\mathrm{d}x\,,\quad\forall\, v\in H_0^1(D), \tag{5b}$$

$$y_\varepsilon-\xi\in H_0^1(D). \tag{5c}$$

For $\varepsilon>0$ small, the penalty term with the $1/\varepsilon$ in (5b) forces the state y_ε to be close to 0 outside Ω_g (for precise, rigorous versions of this statement see

[5, Theorem 2], [8, Theorem 2.2,Theorem 3.1]). It is noted that, even though the above-described method encodes Ω_g and O_p, in fact, as level sets of the functions g, p, respectively, our method is essentially different from the well-known level set method of [6], since no time dependence of the functions g, p and no time evolution of the corresponding open sets Ω_g, O_p is assumed. In particular, we do not need to solve any Hamilton-Jacobi equations in the process.

We now consider a triangular finite element partition of $\overline{D}$, $\overline{D} = \bigcup_{T_h \in \mathcal{T}_h} T_h$, $h > 0$, assuming the grid in D, restricted to E, provides a finite element mesh in E as well. Let V_h, $\widetilde{V}_h$ denote the corresponding finite element spaces in D constructed with piecewise affine continuous functions (with 0 trace on ∂D for elements of V_h). Defining

$$U_{\mathrm{ad}}^h := \big\{(g,p) \in \widetilde{V}_h \times \widetilde{V}_h : g \geq p \text{ on } D \text{ and } p \geq 0 \text{ on } E\big\}, \tag{6}$$

the discretized form of (5) reads

$$\min_{(g_h,p_h)\in U_{\mathrm{ad}}^h} j(g_h,p_h) := \frac{1}{2}\int_E |y_{\varepsilon,h} - y_{d,h}|^2 \,\mathrm{d}x + \frac{1}{2}\int_E |\nabla y_{\varepsilon,h} - \nabla y_{d,h}|^2 \,\mathrm{d}x\,, \tag{7a}$$

$$\begin{aligned}
&\int_D \big[a_1\, H_\varepsilon(p_h) + a_2(1 - H_\varepsilon(p_h))\big]\nabla y_{\varepsilon,h} \cdot \nabla v_h \,\mathrm{d}x \\
&\quad + \int_D \big[b_1\, H_\varepsilon(p_h) + b_2(1 - H_\varepsilon(p_h))\big]\, y_{\varepsilon,h} v_h \,\mathrm{d}x \\
&\quad + \frac{1}{\varepsilon}\int_D \big(1 - H_\varepsilon(g_h)\big)\, y_{\varepsilon,h} v_h \,\mathrm{d}x = \int_D f_h v_h \,\mathrm{d}x\,, \quad \forall\, v_h \in V_h \subseteq H_0^1(D),
\end{aligned} \tag{7b}$$

$$y_{\varepsilon,h} - \xi_h \in V_h \subseteq H_0^1(D), \tag{7c}$$

where (7b) constitutes the equation for the discretized state $y_{\varepsilon,h} \in \widetilde{V}_h \subseteq H^1(D)$, $f_h \in \widetilde{V}_h$ and $\xi_h \in \widetilde{V}_h$ are given suitable approximations of f and ξ, respectively, $g_h, p_h \in \widetilde{V}_h$ are discretized shape functions corresponding to discretizations of Ω_g and O_p, respectively, and $y_{d,h}$ is a suitable given continuous and piecewise affine approximation of y_d.

Similar to [5, Proposition 2] and [8, Corollary 5.3], one obtains the directional derivative of the cost functional $(g,p) \mapsto j(g,p)$ with j as in (7a) at $(g_h,p_h) \in \widetilde{V}_h \times \widetilde{V}_h$ in the direction $(w_h,u_h) \in \widetilde{V}_h \times \widetilde{V}_h$ as

$$\begin{aligned}
\frac{1}{\varepsilon}\int_D H_\varepsilon'(g_h) w_h y_{\varepsilon,h} q_{\varepsilon,h} \,\mathrm{d}x &- \int_D (b_1 - b_2) H_\varepsilon'(p_h) u_h y_{\varepsilon,h} q_{\varepsilon,h} \,\mathrm{d}x \\
&- \int_D (a_1 - a_2) H_\varepsilon'(p_h) u_h \nabla y_{\varepsilon,h} \nabla q_{\varepsilon,h} \,\mathrm{d}x\,,
\end{aligned} \tag{8}$$

where $q_{\varepsilon,h} \in V_h \subseteq H_0^1(D)$ is the solution to the adjoint equation

$$\int_D \big[a_1\, H_\varepsilon(p_h) + a_2(1 - H_\varepsilon(p_h))\big] \nabla q_{\varepsilon,h} \cdot \nabla v_h \,\mathrm{d}x$$
$$+ \int_D \big[b_1\, H_\varepsilon(p_h) + b_2(1 - H_\varepsilon(p_h))\big]\, q_{\varepsilon,h} v_h \,\mathrm{d}x + \frac{1}{\varepsilon} \int_D \big(1 - H_\varepsilon(g_h)\big)\, q_{\varepsilon,h} v_h \,\mathrm{d}x$$
$$= \int_E (y_{\varepsilon,h} - y_{d,h}) v_h \,\mathrm{d}x + \sigma \int_E (\nabla y_{\varepsilon,h} - \nabla y_{d,h}) \cdot \nabla v_h \,\mathrm{d}x\,, \quad \forall\, v_h \in V_h, \tag{9}$$

and the direction of steepest descent $(w_{\mathrm{d},0}, u_{\mathrm{d},0})$ is given by

$$\begin{aligned} w_{\mathrm{d},0} &:= -\,(1/\varepsilon)\, H'_\varepsilon(g_h)\, y_{\varepsilon,h} q_{\varepsilon,h}, \\ u_{\mathrm{d},0} &:= H'_\varepsilon(p_h)\,(a_1 - a_2)\, \nabla y_{\varepsilon,h} \cdot \nabla q_{\varepsilon,h} + H'_\varepsilon(p_h)\,(b_1 - b_2)\, y_{\varepsilon,h} q_{\varepsilon,h}. \end{aligned} \tag{10}$$

While (10) is difficult to use in practise as $H'_\varepsilon(g_h)$ and $H'_\varepsilon(p_h)$ are typically nonzero only in a small neighborhood of $\partial\Omega_{g_h}$ and ∂O_{p_h}, respectively, multiplication by nonnegative coefficients yields the following alternative descent directions $(w_{\mathrm{d},1}, u_{\mathrm{d},1})$ and $(w_{\mathrm{d},2}, u_{\mathrm{d},2})$, without such support restrictions:

$$w_{\mathrm{d},1} := -y_{\varepsilon,h} q_{\varepsilon,h}, \quad u_{\mathrm{d},1} := (a_1 - a_2)\, \nabla y_{\varepsilon,h} \cdot \nabla q_{\varepsilon,h} + (b_1 - b_2)\, y_{\varepsilon,h} q_{\varepsilon,h}, \tag{11}$$
$$w_{\mathrm{d},2} := w_{\mathrm{d},1}\, \chi_S, \qquad u_{\mathrm{d},2} := u_{\mathrm{d},1}\, \chi_S, \tag{12}$$

where χ_S denotes the characteristic function of

$$S := \{x \in D :\, w_{\mathrm{d},1}(x) \geq u_{\mathrm{d},1}(x)\} \cup \{x \in E :\, w_{\mathrm{d},1}(x) \geq 0 \text{ and } u_{\mathrm{d},1}(x) \geq 0\}. \tag{13}$$

Using (12) has the advantage of maintaining the conditions $g \geq p$ on D and $p \geq 0$ on E.

For the numerical results presented below, we employ four variants of an algorithm of gradient with projection type making use of (approximations of) the descent directions (11) and (12). The two variants based on (11) will be called A1a and A1b, whereas the variants based on (12) will be called A2a and A2b. Moreover, variants A1a and A2a will use the admissible set U^h_{ad} of (6), whereas A1b and A2b will use the modification

$$U^h_{\mathrm{ad,b}} := \big\{(g,p) \in U^h_{\mathrm{ad}} :\, |\nabla g|, |\nabla p| \leq 1\big\}, \tag{14}$$

enforcing uniformly bounded gradients for the shape functions, a condition suggested by the results of [8, Sect. 5]. Variant A1a was previously considered in [8]; the remaining three variants are new.

The four algorithms are formulated below in Sect. 2.1, with a description of their implementation in Sect. 2.2. Numerical experiments comparing the performance of the four variants are then presented in Sect. 3.

2 Numerical Algorithms

2.1 Formulation

In preparation for the numerical experiments of Sect. 3, we formulate the employed algorithms. As indicated at the end of the Introduction, we use four variants of

the algorithm previously published in [8], built on the earlier version of [5]. As mentioned above, we denote the four variants by A1a, A1b, A2a, A2b, where A1a is precisely the algorithm used in [8]. A1a and A1b use the descent direction (11) for line searches followed by a projection step, whereas A2a and A2b use (12), which has the advantage of remaining within U_{ad}^h during the line search, avoiding the projection. Variants A1b and A2b project into the smaller space $U_{\mathrm{ad,b}}^h$ of (14) after each line search. The algorithms consist of the following Steps **(1)**–**(7)**:

(1): Set $n := 0$ and choose initial shape functions $(g_{h,0}, p_{h,0}) \in U_{\mathrm{ad}}^h$.

(2): Compute the solution to the state equation $y_n := \theta_{\varepsilon,h}(g_{h,n}, p_{h,n})$, where $\theta_{\varepsilon,h} : \widetilde{V}_h \times \widetilde{V}_h \to \widetilde{V}_h$ denotes the control-to-state operator corresponding to (7b), (7c); and compute the solution to the corresponding adjoint equation $q_n := \tilde{\theta}_{\varepsilon,h}(y_n)$, where $\tilde{\theta}_{\varepsilon,h} : \widetilde{V}_h \to V_h$, $y_{\varepsilon,h} \mapsto q_{\varepsilon,h}$, denotes the solution operator corresponding to (9).

(3): Compute the descent direction $(w_{\mathrm{d}}^n, u_{\mathrm{d}}^n)$, where $w_{\mathrm{d}}^n = w_{\mathrm{d},1}(y_n, q_n)$ and $u_{\mathrm{d}}^n = u_{\mathrm{d},1}(y_n, q_n)$ according to (11) for A1a and A1b, whereas $w_{\mathrm{d}}^n = w_{\mathrm{d},2}(y_n, q_n)$ and $u_{\mathrm{d}}^n = u_{\mathrm{d},2}(y_n, q_n)$ according to (12) for A2a and A2b.

(4): Set $\tilde{g}_{h,n} := g_{h,n} + \lambda_n\, w_{\mathrm{d}}^n$ and $\tilde{p}_{h,n} := p_{h,n} + \lambda_n\, u_{\mathrm{d}}^n$, where $\lambda_n \geq 0$ is determined via line search, i.e. as a solution to the minimization problem

$$\min_{\lambda \geq 0} \; j(g_{h,n} + \lambda\, w_{\mathrm{d}}^n,\, p_{h,n} + \lambda\, u_{\mathrm{d}}^n). \tag{15}$$

(5): For A2a and A2b, set $(\tilde{\tilde{g}}_{h,n}, \tilde{\tilde{p}}_{h,n}) := (\tilde{g}_{h,n}, \tilde{p}_{h,n})$ (no projection is necessary to obtain $(\tilde{\tilde{g}}_{h,n}, \tilde{\tilde{p}}_{h,n}) \in U_{\mathrm{ad}}^h$); for A1a and A1b, set $(\tilde{\tilde{g}}_{h,n}, \tilde{\tilde{p}}_{h,n}) := \pi_h(\tilde{g}_{h,n}, \tilde{p}_{h,n})$, where π_h denotes the projection $\pi_h : \widetilde{V}_h \times \widetilde{V}_h \to U_{\mathrm{ad}}^h$, obtained by first setting $\bar{\tilde{g}}_{h,n}(x_i^h) := \max\{0,\, \tilde{g}_{h,n}(x_i^h)\}$ and $\bar{p}_{h,n}(x_i^h) := \max\{0,\, \tilde{p}_{h,n}(x_i^h)\}$ for each node x_i^h of the triangulation $\mathcal{T}_h$ such that $x_i^h \in \overline{E}$, and second setting $\tilde{\bar{p}}_{h,n}(x_i^h) := \min\{\bar{p}_{h,n}(x_i^h),\, \tilde{\bar{g}}_{h,n}(x_i^h)\}$ for every node x_i^h of the triangulation $\mathcal{T}_h$.

(6): For A1a and A2a, set $(g_{h,n+1}, p_{h,n+1}) := (\tilde{\tilde{g}}_{h,n}, \tilde{\tilde{p}}_{h,n})$ (no second projection necessary); for A1b and A2b, set $(g_{h,n+1}, p_{h,n+1}) := \pi_{h,\mathrm{b}}(\tilde{\tilde{g}}_{h,n}, \tilde{\tilde{p}}_{h,n})$, where $\pi_{h,\mathrm{b}}$ denotes the projection $\pi_{h,\mathrm{b}} : U_{\mathrm{ad}}^h \to U_{\mathrm{ad,b}}^h$, obtained by dividing $\tilde{\tilde{g}}_{h,n}$ and $\tilde{\tilde{p}}_{h,n}$ by α, defined as the max of the max-norms of $|\nabla g|$ and $|\nabla p|$, in case $\alpha > 1$.

(7): RETURN $(g_{h,\mathrm{fin}}, p_{h,\mathrm{fin}}) := (g_{h,n+1}, p_{h,n+1})$ if the change of g, p and/or the change of $j(g, p)$ are below some prescribed tolerance parameter. Otherwise: Increment n, i.e. $n := n + 1$ and GO TO **(2)**.

For all the numerical examples discussed below, we stopped the iteration and returned $(g_{h,\mathrm{fin}}, p_{h,\mathrm{fin}}) := (g_{h,n+1}, p_{h,n+1})$ if $|j(g_{h,n}, p_{h,n}) - j(g_{h,n+1}, p_{h,n+1})| < 10^{-5}$ AND $\|g_{h,n} - g_{h,n+1}\|_2 < 10^{-3}$ AND $\|p_{h,n} - p_{h,n+1}\|_2 < 10^{-3}$, where $|j(g_{h,n}, p_{h,n}) - j(g_{h,n+1}, p_{h,n+1})|/|j(g_{h,n+1}, p_{h,n+1})|$ is used for $|j(g_{h,n}, p_{h,n}) - j(g_{h,n+1}, p_{h,n+1})|$ if $|j(g_{h,n+1}, p_{h,n+1})| > 1$ and analogous for $g_{h,n}$ and $p_{h,n}$.

2.2 Implementation

The state equations as well as the adjoint equations that need to be solved numerically during the above algorithms are discretized linear elliptic PDE with

Dirichlet boundary conditions. The numerical solution is obtained via a finite volume scheme [7, Sect. 4]. More precisely, the software *WIAS-HiTNIHS*[1], originally designed for the solution of more general PDE occurring when modeling conductive-radiative heat transfer and electromagnetic heating [2], has been adapted for use in the present context. *WIAS-HiTNIHS* is based on the program package *pdelib* [1], it employs the grid generator *Triangle* [11] to produce constrained Delaunay triangulations of the domains, and it uses the sparse matrix solver *GSPAR* [3] to solve the linear system arising from the finite volume scheme.

The numerical scheme yields discrete y_n and q_n (cf. Step **(2)** of the above algorithms), defined at each vertex of the triangular discrete grid, interpolated piecewise affine, i.e. affinely to each triangle of the discrete grid. In consequence, the shape functions $g_{h,n}$ and $p_{h,n}$ are piecewise affine as well. Where integrals of these piecewise affine functions need to be computed (e.g. in Step **(7)** of the algorithms), they are computed exactly. A golden section search [10, Sect. 10.2] is used to numerically carry out the minimization (15). Note that the minimization (15) is typically nonconvex and the golden section search will, in general, only provide a *local* min λ_n.

For some numerical examples, the stated initial shape functions $(g_{h,0}, p_{h,0})$ are merely piecewise continuous (cf. the Introduction and [9]) and, thus, not in U^h_{ad}. However, the stated $(g_{h,0}, p_{h,0})$ are only used to determine the values $g_h(x^h_i)$, $p_h(x^h_i)$, at the nodes x^h_i of the triangulation $\mathcal{T}_h$, and the resulting affinely interpolated functions are in U^h_{ad}. Moreover, in Step **(3)** of the algorithms, approximations of the descent directions are used, as for the gradients nodewise averages are computed, that are then affinely interpolated, and the conditions of (13) are enforced nodewise and affinely interpolated. In principle, in might occur that the approximated direction is no longer a descent direction, but such a case was not observed during our numerical experiments.

3 Numerical Experiments

3.1 Numerical Experiments with Precomputed Optimum

The numerical computations of the present section employ the circular fixed domain $D := \left\{(x_1, x_2) : x_1^2 + x_2^2 < 1\right\} \subseteq \mathbb{R}^2$ with fixed subdomain $E := \{(x_1, x_2) \in D : |x_1| > \frac{3}{4},\ |x_2| < \frac{1}{2}\} \subseteq D$ (note E has two connected components). We use a fixed triangular grid provided by *Triangle* [11], consisting of 24458 triangles. The used regularization parameter is $\varepsilon = 10^{-5}$ (cf. [8,9]). The settings for the remaining given quantities are $a_1 := 1$, $a_2 := 10$, $b_1 := 1$, $b_2 := 10$, $f(x_1, x_2) := 5$, $\xi(x_1, x_2) := 2$. The cost functional j as in (7a) depends on the given function $y_{d,h}$. For the first set of numerical results, we precompute $y_{d,h} := y_{\varepsilon,h}$ numerically as the solution to the state Eq. (7b), (7c), using

$$g_h(x_1, x_2) := \begin{cases} -1 & \text{if } (x_1, x_2) \notin E \text{ and } \|(x_1, x_2) - (-1, 0)\|_2 < 0.4, \\ -1 & \text{if } (x_1, x_2) \notin E \text{ and } \|(x_1, x_2) - (1, 0)\|_2 < 0.4, \\ 1 & \text{otherwise,} \end{cases} \tag{16a}$$

[1] **Hi**gh **T**emperature **N**umerical **I**nduction **H**eating **S**imulator.

$$p_h(x_1, x_2) := \begin{cases} 1 & \text{in } E, \\ -1 & \text{in } D \setminus E. \end{cases} \tag{16b}$$

The computed $y_{d,h}$ with the corresponding Ω_g and O_p is depicted in Fig. 1. Using the precomputed $y_{d,h}$ has the advantage that we actually know $y_{d,h}$ together with g_h, p_h as in (16) provides an absolute minimum in the following numerical examples, employing the cost functional j of (7a) with the precomputed $y_{d,h}$ from above. A series of four numerical experiments was conducted, all using the initial shape functions $g_{h,0}(x_1, x_2) := 1$, $p_{h,0}(x_1, x_2) := 1$ (see Fig. 2).

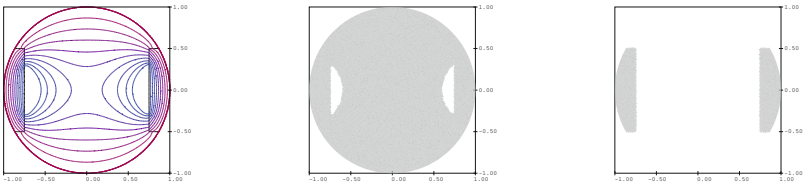


Fig. 1. Precomputed $y_{d,h}$ used in all experiments of Sect. 3.1 (left, isolevels spaced at 0.2), obtained as the solution to the state Eq. (7b), (7c); with the corresponding Ω_g (middle) and O_p (right).

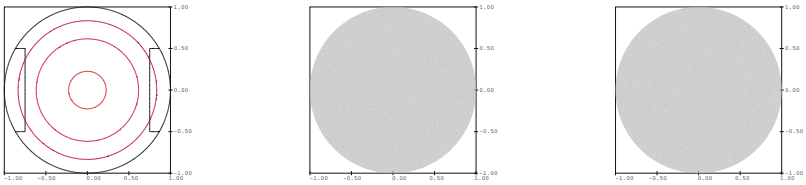


Fig. 2. Initial state, shapes used in all experiments of Sect. 3.1. Left: State isolevels spaced at 0.2. Middle: Shape Ω_g. Right: Shape O_p. Cost: $j(g_{h,0}, p_{h,0}) = 19.0$..

We refer to the experiments as 1:A1a, 1:A1b, 1:A2a, and 1:A2b, depending on which variant of the algorithm of Sect. 2.1 was used. The results for 1:A2a and 1:A2b are shown in Fig. 3. The final state and shapes for 1:A1a and 1:A1b were very similar to those of 1:A2a, with slightly higher final costs (0.69 and 0.28, respectively). All variants reduce the cost significantly, all resulting local minima being different and different from the absolute min. Variant A2a gives the best result, whereas A2b results in the highest final cost, where one also observes a symmetry breaking due to the discrete grid. Actually, for A2b, after the first line search, the cost is 0.29 with shapes resembling the final shapes of the other variants, but the projection of Step **(6)** can subsequently result in a cost increase, which occurs in this example.

3.2 Numerical Experiments Without Precomputed Optimum

In contrast to the experiments of the previous section, we now consider a setting, where we are no longer in the situation of a known precomputed optimum. For

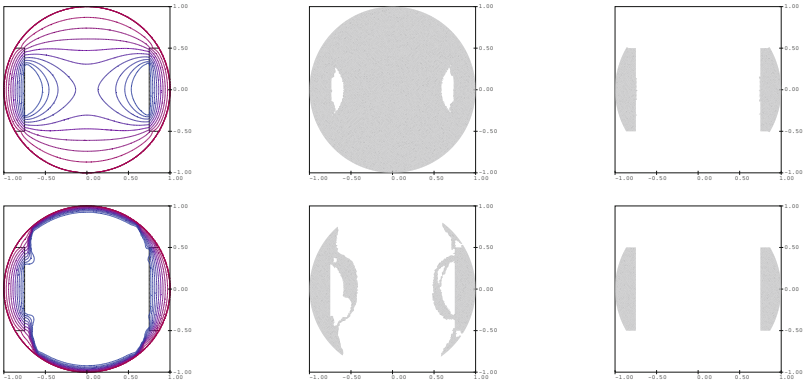


Fig. 3. Final state, shapes for shape optimizations 1:A2a (1st row) and 1:A2b (2nd row) of Sect. 3.1. Left: State isolevels spaced at 0.2. Middle: Shapes Ω_g. Right: Shapes O_p. Final costs $j(g_{h,\mathrm{fin}}, p_{h,\mathrm{fin}})$ are 0.053 for 1:A2a, 1.30 for 1:A2b. Required number of line searches: 6 for 1:A2a, 29 for 1:A2b

the following numerical results, the fixed domain D is still the unit disk as in Sect. 3.1. However, the fixed subdomain E is now at the bottom of D, defined by $E := \{(x_1, x_2) \in D : x_2 < -0.7\} \subseteq D$. The numerical computations employ a fixed triangular grid provided by *Triangle* [11], consisting of 24623 triangles. The parameter settings are as in Sect. 3.1, except for $f(x_1, x_2) := 10(x_1^2 + x_2^2) + 5$. The cost functional is as in (7a) with $y_{d,h}(x_1, x_2) := x_1 + x_2$. A series of four numerical experiments was conducted, all using the initial shape functions $g_{h,0}(x_1, x_2) := p_{h,0}(x_1, x_2) := \begin{cases} 1 & \text{if } (x_1, x_2) \in E, \\ -1 & \text{otherwise} \end{cases}$ (see Fig. 4).

We refer to the experiments as 2:A1a, 2:A1b, 2:A2a, and 2:A2b, depending on which variant of the algorithm of Sect. 2.1 was used. Results are shown in Fig. 5, except for 2:A2a, which converged after 10 line searches to a local min almost identical to the initial condition. All other variants reduce the cost significantly, all resulting local minima being different. Here, the lowest final cost is achieved for variant A1b. One notices significant changes in shapes (including topology changes) during the optimizations, where very different shapes can result in nearly identical costs. As in Sect. 3.1, symmetry breaking can occur due to the discrete grid.

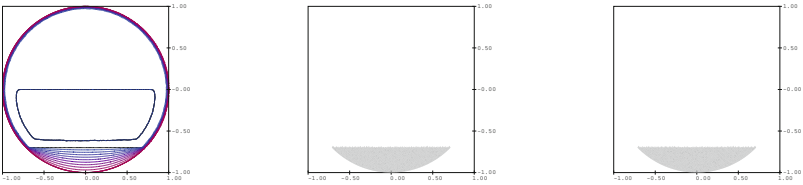


Fig. 4. Initial state, shapes used in all experiments of Sect. 3.2. Left: State isolevels spaced at 0.2. Middle: Shape Ω_g. Right: Shape O_p. Cost: $j(g_{h,0}, p_{h,0}) = 24.8$.

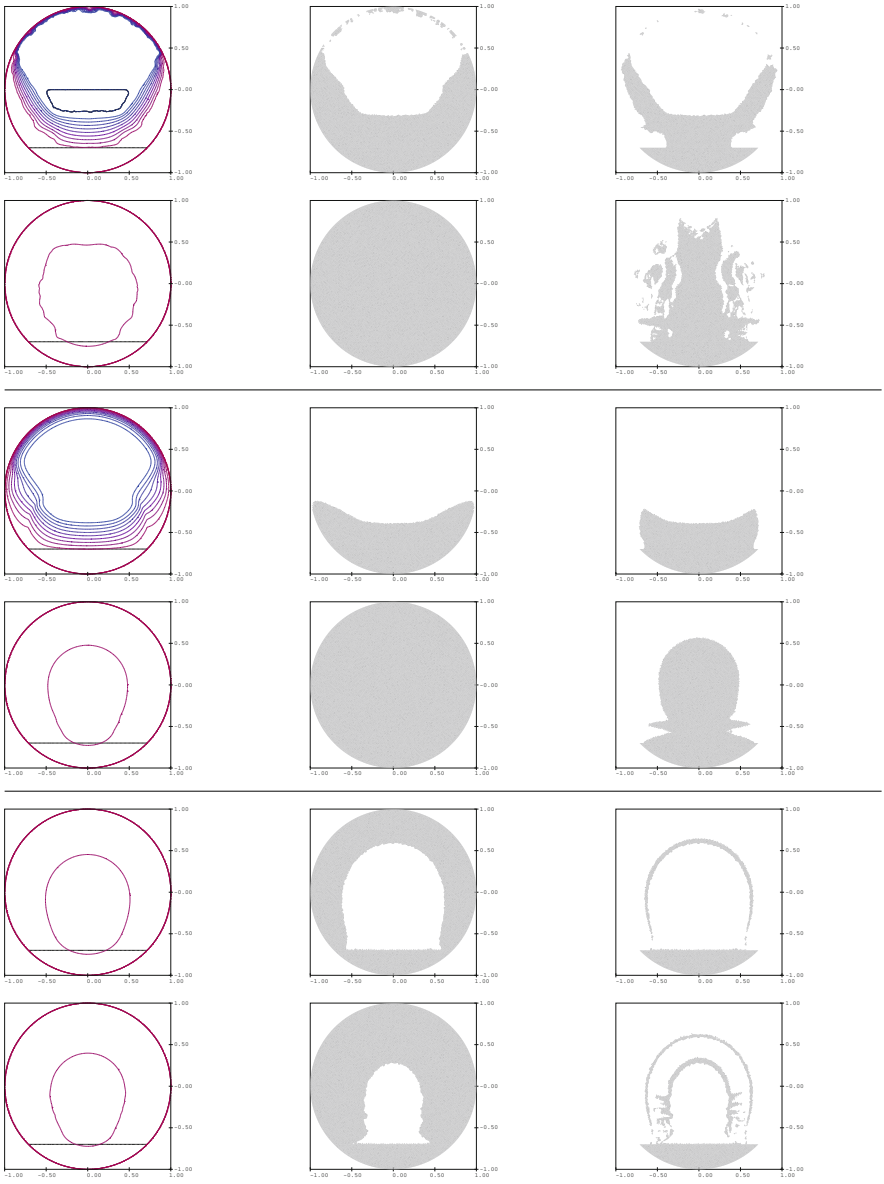


Fig. 5. Intermediate and final state, shapes for shape optimizations of Sect. 3.2, i.e. for Experiments 2:A1a (1st, 2nd row), 2:A1b (3rd, 4th row), and 2:A2b (5th, 6th row). Left: State isolevels spaced at 0.2. Middle: Shapes Ω_g. Right: Shapes O_p. Costs at shown intermediate states are 2.64 for 2:A1a, 1.74 for 2:A1b, 1.70 for 2:A2b. Final costs $j(g_{h,\mathrm{fin}}, p_{h,\mathrm{fin}})$ are 1.72 for 2:A1a, 1.68 for 2:A1b, 1.69 for 2:A2b. Number of line searches for intermediate and for final state: 8 and 31 for 2:A1a, 2 and 29 for 2:A1b, 3 and 38 for 2:A2b

4 Conclusions

In a series of numerical experiments, we have studied four variants of an algorithm of gradient with projection type for shape optimization problems driven by elliptic PDE. The variants used different descent directions and different sets of admissible shape functions. Except in one situation, all variants were effective in finding local minima of significantly reduced costs. However, it did depend on both the equation and on the initial condition, which variant showed the best performance. Thus, further research seems warranted to further evaluate and improve the different variants.

References

1. Fuhrmann, J., Koprucki, T., Langmach, H.: pdelib: an open modular tool box for the numerical solution of partial differential equations. Design patterns. In: Proceedings of the 14th GAMM Seminar on Concepts of Numerical Software, Kiel, 23–25 January 1998, University of Kiel, Kiel, Germany (2001)
2. Geiser, J., Klein, O., Philip, P.: Numerical simulation of temperature fields during the sublimation growth of SiC single crystals, using WIAS-HiTNIHS. J. Cryst. Growth **303**, 352–356 (2007)
3. Grund, F.: Direct linear solvers for vector and parallel computers. In: Hernández, V., Palma, J.M.L.M., Dongarra, J. (eds.) VECPAR 1998. LNCS, vol. 1573, pp. 114–127. Springer, Heidelberg (1999)
4. Mäkinen, R., Neittaanmäki, P., Tiba, D.: On a fixed domain approach for a shape optimization problem. In: Ames, W., van Houwen, P. (eds.) Computational and Applied Mathematics II: Differential Equations, North Holland, Amsterdam, pp. 317–326 (1992)
5. Neittaanmäki, P., Pennanen, A., Tiba, D.: Fixed domain approaches in shape optimization problems with Dirichlet boundary conditions. Inverse Prob. **25**(5), 1–18 (2009)
6. Osher, S., Sethian, J.: Fronts propagating with curvature dependent speed: slgorithms based on Hamilton-Jacobi formulations. J. Comput. Phys. **79**(1), 12–49 (1988)
7. Philip, P.: Analysis, optimal control, and simulation of conductive-radiative heat transfer. Math. Appl./Ann. AOSR **2**, 171–204 (2010)
8. Philip, P., Tiba, D.: A penalization and regularization technique in shape optimization problems. SIAM J. Control Optim. **51**(6), 4295–4317 (2013)
9. Philip, P., Tiba, D.: Shape optimization via control of a shape function on a fixed domain: theory and numerical results. In: Repin, S., Tiihonen, T., Tuovinen, T. (eds.) Numerical Methods for Differential Equations, Optimization, and Technological Problems, Computational Methods in Applied Sciences, vol. 27, pp. 305–320. Springer, New York (2013)
10. Press, W., Teukolsky, S., Vetterling, W., Flannery, B.: Numerical Recipes: The Art of Scientific Computing, 3rd edn. Cambridge University Press, New York (2007)
11. Shewchuk, J.: Delaunay refinement algorithms for triangular mesh generation. Comput. Geom. **22**(1–3), 21–74 (2002)
12. Woo, H., Kim, S., Seol, J., Lionheart, W., Woo, E.: A direct tracking method for a grounded conductor inside a pipeline from capacitance measurements. Inverse Prob. **22**, 481–494 (2006)

Tomographic Reconstruction of Homogeneous 2D Geometric Models with Unknown Attenuation

Zenith Purisha(✉) and Samuli Siltanen

Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
{zenith.purisha,samuli.siltanen}@helsinki.fi

Abstract. A new method is presented for tomographic reconstruction of objects with homogeneous attenuation. The method is based on parametric representation with Non-Uniform Rational B-Splines (NURBS) and statistical inversion with a Markov Chain Monte Carlo (MCMC) algorithm. The method recovers the approximate boundary curve shape and the attenuation value of two-dimensional homogeneous objects. The boundary can be represented by NURBS with few parameters, reducing the number of degrees of freedom. However, this leads to a nonlinear inverse problem, and therefore statistical inversion is used. One of the benefits of the approach is that the reconstruction is automatically in the form of the geometrical representation in industrial CAD format or CNC configuration. Computational results are presented with two different simulated homogeneous geometric models and sparsely sampled tomographic data. The new method outperforms the baseline method (filtered back-projection) in image quality but not in computational speed.

Keywords: Tomography · Homogeneous · CAD · NURBS · Bayesian inversion · MCMC

1 Introduction

Creating a virtual model of a given physical object is increasingly important in, for example, reverse engineering and game development. The details of reconstructing the model depend on the kind of measurements that are available about the object. For example laser scanning and digital photography are popular methods providing surface information. In this work we concentrate on sparsely sampled X-ray tomography measurements.

Consider a three-dimensional cylindrical object $\Omega \times \mathbb{R}$ with a simply connected base $\Omega \subset \mathbb{R}^2$. Furthermore, assume that we only know that the object is homogeneous: the X-ray attenuation coefficient has an unknown but constant value $c > 0$ inside the object.

This work was supported by the Academy of Finland through the Finnish Centre of Excellence in Inverse Problems Research 2012–2017, decision number 250215.

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 247–256, 2014.
DOI: 10.1007/978-3-662-45504-3_24

We discuss the situation where we have X-ray projection data of a transversal slice of the object. In other words, we have access to a collection of line integrals of the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x,y) = \begin{cases} c & \text{for } (x,y) \in \Omega, \\ 0 & \text{for } (x,y) \in \mathbb{R}^2 \setminus \Omega. \end{cases} \tag{1}$$

The angular sampling of the X-ray data can be very sparse, allowing for quick measurement process with low radiation dose. Our aim is to recover two things: the boundary $\partial\Omega \subset \mathbb{R}^2$ represented as a parameterized curve and the attenuation coefficient c.

We want our method to be practically useful in industrial environments. Computer numerical control (CNC) machines are widely used in modern production facilities, and they use computer-aided design (CAD) models. The proposed tomographic method represents the unknown boundary curve in Non-Uniform Rational Basis Spline (NURBS) form, which is the standard in CAD software. This direct connection with industrial standards is the main motivation behind the proposed method.

NURBS curves are represented by a relatively small number of parameters: a set of planar *control points* and a related *knot vector*. In this paper we fix the knot vector, so the information to be recovered consists merely of the control points and the attenuation parameter. The low dimensionality of this problem formulation offers computational advantages. However, there is a complication as well: the linear inverse problem of X-ray tomography becomes nonlinear in this parameterization. Therefore, we resort to the very general framework of Bayesian inversion [4,10].

In Bayesian inversion, limited measurement data is complemented by *a priori* information using the Bayes formula. This way the ill-posed inverse problem is recast in a well-posed form of exploring the posterior probability distribution. As explained in [5], in the case of X-ray tomography this involves a discrete attenuation model and a Monte Carlo Markov Chain (MCMC) method for sampling the posterior. Usually the large number of pixels in the reconstructed image leads to MCMC sampling in a very high-dimensional space (one dimension for each pixel). In our case the posterior distribution is defined in a relatively low-dimensional space: one dimension for the attenuation value plus two dimensions for each control point. This enables efficient MCMC sampling.

The *a priori* information we use is rather simple: we assume that we have an upper bound for the diameter of the two-dimensional shape (transversal slice) under measurement. Also, we assume that the curve does not have too small details (parts with very high curvature) and choose the number of control points to be as small as possible while still capable of representing the smallest details in the curve.

We demonstrate the novel NURBS-MCMC method using two simulated non-convex examples. See Fig. 3 below. The reconstruction algorithm is found to recover the attenuation coefficient quite precisely and the boundary shape with reasonable accuracy from very sparsely sampled X-ray data (only 18 projection directions).

This paper is organized as follows. In Sect. 2, we discuss the theory of NURBS curves. In Sect. 3 we present the X-ray measurement model. Section 4 is devoted to the description of Bayesian inversion. In Sect. 5, we present the reconstruction results, and in Sect. 6 we conclude our findings.

2 NURBS Description for Parametric Curve

We model an unknown object boundary $\partial\Omega$ by a continuous curve $\mathcal{S} : [0,1] \to \mathbb{R}^2$. In our computational problem, we construct $\mathcal{S}$ using NURBS that are widely used as computationally fast and robust representations of curves.

The basic building blocks of NURBS are the following:

1. *Control points* $\mathbf{p}_1, \ldots, \mathbf{p}_n$. These planar locations $\mathbf{p}_i \in \mathbb{R}^2$ are, roughly speaking, points of attraction for the NURBS curve, where $i = 1, 2, ..., n$. Throughout the paper we denote by n the number of control points.
2. *Knots* $t_1, t_2, \ldots, t_K \in [0,1]$, with ordered as follows:

$$0 = t_1 \leq t_2 \leq \ldots \leq t_K = 1,$$

where $K > n$. The knot are used to divide the interval $[0,1]$ into suitable pieces. We collect the knots into a knot vector $[t_1 \quad t_2 \quad \ldots \quad t_K]$.
3. *Basis function* $(N_{i,p}(t))$ specifies how strongly the control point $\mathbf{p}_i$ attracts the NURBS curve. The first-order basis function is

$$N_{i,1}(t) = \begin{cases} 1 & \text{if } t_i \leq t < t_{i+1}, \\ 0 & \text{otherwise.} \end{cases}$$

Higher-order basis functions are defined recursively as

$$N_{i,p}(t) = \frac{t - t_i}{t_{i+p} - t_i} N_{i,p-1}(t) + \frac{t_{i+p+1} - t}{t_{i+p+1} - t_{i+1,p-1}} N_{1+i,p-1}(t),$$

where p is the order of the basis function and $K = n + p$.
The general form of NURBS curve can be written as

$$\mathcal{S}(t) = \sum_{i=1}^{n} \mathbf{p}_i R_{i,p}(t). \tag{2}$$

The *Rational* in NURBS comes from the rational function $R_{i,p}(t) = \frac{\omega_i N_{i,p}(t)}{\sum_{i=0}^{n} \omega_i N_{i,p}(t)}$, where the weights $\omega_i \geq 0$, for all i. In this preliminary result, we use the same weights for all control points.

3 Tomographic Measurement Model

Consider a continous tomography model $f : \mathbb{R}^2 \to \mathbb{R}$ as in Eq. (1), where $f(x,y) \geq 0$ and $\text{supp}(f) \subset \Omega$ with bounded $\Omega \subset \mathbb{R}^2$.

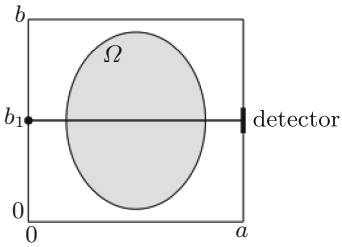


Fig. 1. An X-ray travels along a homogeneous target slice. The shade of gray describes a constant coefficient attenuation c inside Ω.

Consider an X-ray traveling through a two dimensional object along a straight line as shown in the Fig. 1. In this specimen, the slice of the target is in square defined by $0 \leq x \leq a$ and $0 \leq y \leq b$. Assume that an X-ray penetrates along the horizontal path $0 \leq x \leq a$ and $y = b_1$.

Let us consider that the X-ray has the initial intensity $I_0 = I(0)$ and the intensity becomes smaller, say $I_1 = I(1)$ after it passes the object. This situation can be modeled using $f(x, y)$, an attenuation coefficient function, as:

$$\frac{dI(x)}{I(x)} = -f(x, b_1)dx,$$

where $I(x)$ is the intensity of the X-ray at the point (x, b_1) while passing through the source to the detector.

In tomographic imaging, we want to collect information about f using different angles. Let us consider the Radon transform, denoted by $\mathcal{R}$, as follows. Assume $\alpha \in \mathbb{R}$ as an angle measured in radians:

$$\boldsymbol{\alpha} = \begin{bmatrix} \cos\alpha \\ \sin\alpha \end{bmatrix} \in \mathbb{R}^2,$$

the unit vector with angle α with respect to the $x-$axis.

The radon function of the function f depends on the angular parameter α and on a linear parameter $s \in \mathbb{R}$ as follows:

$$\mathcal{R}f(s, \alpha) = \int_{\mathbf{x}.\boldsymbol{\alpha}=s} f(\mathbf{x})d\mathbf{x}^{\perp},$$

where $d\mathbf{x}^{\perp}$ is the one dimensional Lebesgue measure along the line $\{\mathbf{x} \in \mathbb{R}^2 : \mathbf{x} \cdot \boldsymbol{\alpha} = s\}$.

For computational reasons, we need a discrete model. In this case, we construct two discrete models: a pixel-based object model and a NURBS-based object model, a model where the boundary $\partial\Omega$ is expressed as a NURBS curve as shown in Fig. 2.

In the pixel-based model, the line integral is discretized using the standard pencil-beam model. We use the pixel-based Matlab routine `radon.m` for simulating parallel-beam tomographic data.

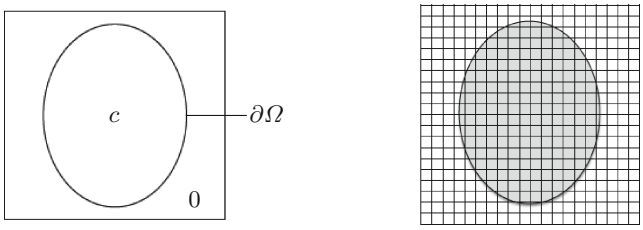


Fig. 2. Left: the NURBS-based object model where $\partial\Omega$ is a NURBS curve. The inside of the curve is set to be c and the outside is set to 0. Right: the pixel image

In the NURBS-based model, the line integral is discretized by moving to pixel-based model using an operator $\mathcal{B}$ defined by

$$\mathcal{B}(\mathbf{p}, c) = \begin{cases} c, & \text{if the pixel center is inside the NURBS curve,} \\ 0, & \text{if the pixel center is outside the NURBS curve.} \end{cases} \tag{3}$$

Assuming that the knot vector is fixed, the degrees of freedom in our NURBS model are the control points $\mathbf{p}_1, \ldots, \mathbf{p}_n$ together with the attenuation.

In the simulation, we measure two simple homogeneous shapes that have different attenuation. To avoid inverse crime [7], we produce the synthetic phantoms Ω_1 and Ω_2 without using NURBS. Those objects are set to be homogeneous inside with attenuation values 2 and 3,5, respectively, as shown in Fig. 3.

The objects are measured with the resolution 64×64 using parallel beam geometry as shown in Fig. 4. From the source, the X-ray penetrates through the objects and a sensor detects the projection images from different directions. Sparse full angle data, $0°, 10°, 20°, ..., 170°$, are applied to obtain the projections and each direction consists of 95 lines.

Fig. 3. Homogeneous phantoms

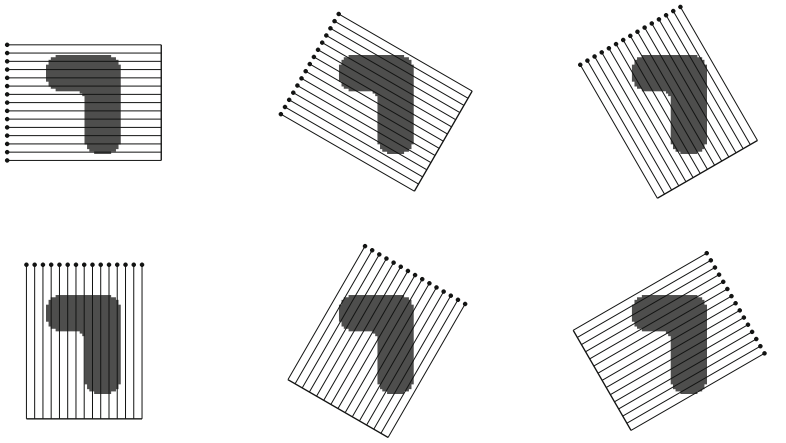


Fig. 4. Parallel beam X-ray measurement geometry. There are 6 different directions $(0°, 30°, 60°, 90°, 120°, 150°)$ and 15 lines. Black dots show the locations of the X-ray source at different times of measurement. The thick line represents the detector measuring the intensity of the X-rays after passing through the target

4 Bayesian Inversion for Control Points and Attenuation Value

This section presents the Bayesian approach to handle the inverse problem. This measurement data, $\mathbf{m}$, is used to get information about other quantities. In this case, we encounter a nonlinear inverse problem, which need to be solved by recovering $\mathcal{B}$ that depends on $\mathbf{p}$ and c.

We model the problem as the following form:

$$\mathbf{m} = \mathcal{R}(\mathcal{B}(\mathbf{p}, \mathbf{c})) + \varepsilon, \tag{4}$$

where ε is the error of the measurement.

The Bayesian inversion approach is based on the relations between probability distributions to model the inadequacy of information in an inverse problem. Before performing the collection of measurement data, we construct a model for *a priori* knowledge. Since the control points are presented in polar coordinates, i.e. $\mathbf{p}_i = (r_i \sin\theta_i, r_i \cos\theta_i)$, we assume that the angle of each parameter is not less than $\theta_i^{\min}$ and not more than $\theta_i^{\max}$, and the distance of each parameter from the central point of the object is nonnegative and not more than $r_i^{\max}$. In this case, Ω_1 and Ω_2 have $r_i^{\max}$ values that equal to 15 and 30, respectively, and the maximum of the attenuation value $c_{\max}$ is 5 for both objects.

We formulate the *prior* condition as follows

$$\pi(\mathbf{m} \,|\, (\mathbf{p}, c)) = \begin{cases} \exp(-\frac{1}{2\sigma_2^2}\|(\mathbf{p}, c) - (\tilde{\mathbf{p}}, \tilde{c})\|_2^2) & \text{for } 0 \le r_i \le r_i^{\max} \text{ and } 0.1 < c < c_{\max} \\ & \text{and } \theta_i^{\min} \le \theta_i \le \theta_i^{\max}, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $\mathbf{p} = \{\mathbf{p}_i\}$ and $(\tilde{\mathbf{p}}, \tilde{c})$ is *a priori* information of the position of control points and the attenuation value. After examining the measurement setting and *prior* information, we can model the conditional probability of $\mathbf{m}$, which is called the *likelihood* function

$$\pi(\mathbf{m} \,|\, (\mathbf{p}, c)) = \exp(-\frac{1}{2\sigma_1^2}\|\mathcal{R}(\mathcal{B}(\mathbf{p}, c)) - \mathbf{m}\|_2^2). \tag{6}$$

By given observed data, $\mathbf{m}$, the conditional probability $\pi(\mathbf{p}, c \,|\, \mathbf{m})$ of $\mathbf{p}$ and c can be expressed as follows

$$\pi(\mathbf{p}, c \,|\, \mathbf{m}) = \frac{\pi(\mathbf{p}, c)\pi(\mathbf{m} \,|\, (\mathbf{p}, c))}{\pi(\mathbf{m})}, \tag{7}$$

which is called the *posterior* distribution. To solve our inverse problem, we need to explore this distribution.

As a common method to represent statistical estimates, we apply the *conditional mean* (CM) of the unknown $\mathbf{p}$ and c. Since CM is defined as

$$(\mathbf{p}^{\mathrm{CM}}, c^{\mathrm{CM}}) = \int_{\mathbb{R}^N} (\mathbf{p}, c)\pi(\mathbf{p}, c \,|\, \mathbf{m})\, d(\mathbf{p}, c),$$

finding the estimate leads to the integration problems. Typically, the integration is over a high-dimensional space. To unfold this issue, a Markov chain Monte Carlo (MCMC) technique is recommended to generate a sample from the *posterior* distribution. For a general introduction to Bayesian inversion and properties of MCMC computation see [2–4].

By applying the CM estimate to the samples $\{\mathbf{p}_{1,l}, \mathbf{p}_{2,l}, ..., \mathbf{p}_{n,l}, c_l\}$, we get

$$\mathbf{p}_i^{\mathrm{CM}} \approx \frac{1}{N}\sum_{l=1}^{N} \mathbf{p}_{i,l}^{\mathrm{CM}} \text{ and } c^{\mathrm{CM}} \approx \frac{1}{N}\sum_{l=1}^{N} c_l^{\mathrm{CM}},$$

where $\mathbf{p}^{\mathrm{CM}} = \{\mathbf{p}_i^{\mathrm{CM}}\}$, $i = 1, 2, ..., n$ and N is the number of evaluations. For the NURBS curve reconstruction with N evaluations, it is written as $\mathcal{S}_N^{\mathrm{CM}}$.

5 Computational Results

In this section, numerical examples are presented. We use Metropolis-Hastings as sampling algorithm to generate control points and attenuation with 1 000 000 iterations (applied also to the Radon transform and its adjoint). In each iteration, the weights are set to be equal while the order and knot vector of NURBS curve are set to be fixed. The order is set to be 3 because it is widely used in practical application and to avoid heavy calculation times. As a default knot vector in CAD, the open uniform knot vector is chosen for Ω_1 and Ω_2, $[0\ 0\ 0\ \frac{1}{11}\ \frac{2}{11}\ \frac{3}{11}\ \frac{4}{11}\ \frac{5}{11}\ \frac{6}{11}\ \frac{7}{11}\ \frac{8}{11}\ \frac{9}{11}\frac{10}{11}\ 1\ 1\ 1]$ and $[0\ 0\ 0\ \frac{1}{7}\ \frac{2}{7}\ \frac{3}{7}\ \frac{4}{7}\ \frac{5}{7}\ \frac{6}{7}\ 1\ 1\ 1]$, respectively.

The NURBS curves as in the rightmost Fig. 5 are achieved. By using the mapping as in (3), both final shape reconstructions are presented in the middle of Figs. 6 and 7. The error in the shape reconstructions is given as follows. Denote O as the image of the original 2D object and O^{rec} as the image of the reconstruction. Set $O \backslash O^{\mathrm{rec}}$ for points that belong to the original object but not to the reconstruction and $O^{\mathrm{rec}} \backslash O$ for points that belong to the reconstruction but not to the original object. The relative error in the reconstruction is written as

$$\frac{(\mathrm{area}(O \backslash^{\mathrm{rec}}) + \mathrm{area}(O^{\mathrm{rec}} \backslash O))}{\mathrm{area}\ O} 100\,\% \tag{8}$$

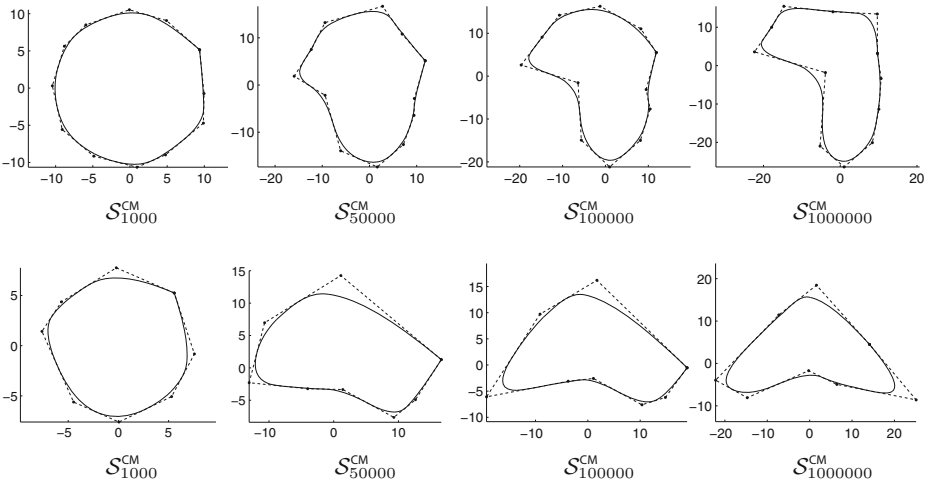


Fig. 5. The thin black line is the target curve. The thick black line is the reconstruction of the NURBS curve, $\mathcal{S}_N^{\mathrm{CM}}$. Top: reconstructions for Ω_1, bottom: reconstructions for Ω_2. The black circle markers are the control points, $\mathbf{p}^{\mathrm{CM}}$

Fig. 6. Left: Original Ω_1. Center: NURBS-MCMC reconstruction. Right: FBP reconstruction. Both are using error 0.1 %

Fig. 7. Left: Original Ω_2. Center: NURBS-MCMC reconstruction. Right: FBP reconstruction. Both are using error 0.1 %

By applying (8), the relative errors of Ω_1 and Ω_2 reconstructions using NURBS-MCMC are 15 % and 8.3 %, respectively. Recovered chains of attenuation values of Ω_1 and Ω_2 after *burn-in* period have relative errors 9.26 % and 1.47 %, respectively.

Table 1. Mean and standard deviation of FBP reconstruction

	mean	standard deviation
Ω_1	1.9795	0.11
Ω_2	3.49	0.14

Table 2. Computation time (in seconds) for all reconstruction methods.

FBP	NURBS-MCMC	
1	18 000	

The rightmost images in Figs. 6 and 7 show recovered shapes using filtered back projection (FBP). The reconstruction uses the resolution 64×64. To assess the error in the reconstructed attenuation value, a representative rectangular region of interest is picked from the inside the reconstruction. The mean and the standard deviation of the recovered attenuation values are computed as we can see in Table 1, while Table 2 shows computation times for both methods.

6 Discussion and Conclusions

Reconstruction using the NURBS-MCMC method in nonlinear inverse problem can recover measurement data successfully. Homogeneous objects Ω_1 and Ω_2 are recovered by only $2n+1$ parameters: 25 and 17, respectively. Those recovered data are geometrical representations which are automatically set to CAD or CNC configuration. In the middle of Figs. 6 and 7, the vector graphic form is converted to be 512×512.

In filtered backprojection, the reconstruction is represented by pixel images and consequently doing a segmentation to represent the shape is nontrivial.

Nevertheless, the slowness of computation is a shortcoming of the proposed method as we can see in Table 2, but by implementing parallel computing, the problem can be handled.

References

1. Bertrand, C., et al.: A probabilistic solution to the MEG inverse problem via MCMC methods: the reversible jump and parallel tempering algorithms. IEEE Trans. Biomed. Eng. **48**(5), 533–542 (2001)
2. Gamerman, D., Lopes, H.F.: Markov Chain Monte Carlo In Practice. Chapman and Hall/CRC, Boca Raton (1996)
3. Gilks, W.R., et al.: Markov Chain Monte Carlo : Stochastic Simulation for Bayesian Inference. Chapman and Hall/CRC, Boca Raton (2006)

4. Kaipio, J.: Statistical and Computational Inverse Problems, vol. 160. Springer, New York (2005)
5. Kolehmainen, V., Siltanen, S., et al.: Statistical inversion for medical x-ray tomography with few radiographs: II. Application to dental radiology. Phys. med. Biol. **48**, 1465–1490 (2003)
6. Marzouk, Y.M., Habib, N.N.: Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems. J. Comput. Phys. **228**(6), 1862–1902 (2009)
7. Mueller, J., Siltanen, S.: Linear and Nonlinear Inverse Problems with Practical Applications. Computational Science and Engineering. SIAM, Philadelphia (2012)
8. Renken, F., Subbaraya, G.: NURBS-based solutions to inverse problems in droplet shape prediction. Comput. Methods Appl. Mech. Eng. **190**, 1391–1406 (2000)
9. Rogers, D.F.: An Introduction to NURBS : with historical perspective, vol. 1. Academic Press, Morgan Kaufmann (2001)
10. Tarantola, A.: Inverse problem theory and methods for model parameter estimation. Society for Industrial and Applied Mathematics, Philadelphia (2005)
11. Wang, J., Zabaras, N.: Hierarchical Bayesian models for inverse problems in heat conduction. Inverse Prob. **21**, 21–183 (2004)

A Control Delay Differential Equations Model of Evolution of Normal and Leukemic Cell Populations Under Treatment

I. Rodica Rădulescu(✉), Doina Cândea, and Andrei Halanay

Department of Mathematics and Informatics, Politehnica University of Bucharest, Splaiul Independentei 313, 060042 Bucharest, Romania
nicola_rodica@yahoo.com

Abstract. The dynamics and evolution of leukemia is determined by the interactions between normal and leukemic cells populations at every phase of the development of hematopoietic cells. For both types of cell populations, two subpopulations are considered, namely the stem-like cell population (i.e. with unlimited self-renew ability) and a more mature, differentiated one, possessing only the capability to undergo limited reproduction. Treatment effects are included in the model as functions of time and a cost functional is considered. The optimal control is obtained using a discretization scheme. Numerical results are discussed in relation to the medical interpretation.

Keywords: Leukemia · Asymmetric division · Competition · Optimal control · Treatment

AMS Classification: 34K35, 37N25, 92C50, 93C23.

1 Introduction

For the description of biological processes implied in hematopoiesis, a mathematical model that includes time delays will be used. It is based on the mass action principle, in the spirit of [1,2,4,12,16,20]. Other authors [14,15] used the more simple model from [18] for the dynamics of hematopoietic stem cells (HSC).

Chronic Myelogenous Leukemia (CML), also known as Chronic Granulocytic Leukemia, is a cancer of white blood cells. It is a clonal marrow stem cells disorder in which the main characteristic is the proliferation of granulocytes (neutrophils, eosinophils and basophils) and of their precursors in the bone marrow and the accumulation of these cells in the blood. It is a type of myeloproliferative disease associated to a chromosomal translocation called the Philadelphia chromosome (see also [6,19]) presenting the oncogene BCR-ABL that encodes a tyrosine kinase protein. Tyrosine kinases are enzymes that play an important role in

This work was supported by CNCS-ROMANIA Grant ID-PCE-2011-3-0198.

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 257–266, 2014.
DOI: 10.1007/978-3-662-45504-3_25

tumor development by supporting cell growth through phosphorylation of signaling proteins [11]. Understanding the molecular mechanism of CML permitted the development of specific tyrosine kinase inhibitors (TKIs) as imatinib (Gleevec), dasatinib or nilotinib. The standard first line therapy is nowadays imatinib, which acts through competitive inhibition at the ATP-binding site of the BCR-ABL enzyme, leading to the inhibition of tyrosine phosphorylation of proteins involved in BCR-ABL signal transduction [7]. The molecular effect of imatinib is mainly the inhibition of cell proliferation of BCR-ABL–positive cells but, there are experimental evidences that, in the mature cell lines, the inhibition of cell proliferation is followed by apoptosis [11]. Although imatinib has a very good successful rate, there are many experimental evidences attesting it does not affect quiescent stem cells deep in the bone marrow, and the consequence is the disease reapers after the treatment is stopped.

In this paper, we study an optimal control delay differential equation model of four cell populations, namely two healthy and two leukemic. For these classes of cells, we consider a population of mature cells which lost their self-renew ability and a population of stem-like cells involving a larger category consisting of proliferating stem and progenitor cells with self-renew capacity. The emphasis in this optimal control model is on establishing treatment strategies, considering the competition of healthy vs. CML cell populations and three types of division that a stem-like cell can exhibit: self-renew, asymmetric division and differentiation [4,17,20,21,24].

Of course, besides a correct mathematical model for the time evolution of the studied cell populations, it is very important to model the treatment effect as accurate as one can. Obviously, different drugs have different effects: some affect not only leukemic populations but also healthy ones (the cytotoxic ones), some kill the cells while others only delay or stop the division process (the cytostatic ones). We take into account here the standard treatment protocol with imatinib which, as we specified, acts by inhibiting the BCR-ABL signal transduction. In this way, it restrains the proliferative advantage of the CML cells and healthy cells regain their advantage. Moreover, it is straightforward to say that imatinib restores most of the abnormal functions of the CML cells, and the most important function affected by the drug is the division process of CML cell population. However, it is uncertain to what extent it affects all three kinds of division and therefore, we consider here the hypothesis that imatinib influences self-renew, asymmetric division and differentiation equally.

2 Description of the Model

In the present paper, is assumed that the hematopoietic stem cells that are considered are in the proliferative phase or spend a short time into the resting phase. These cells are called, following [17], Short-Term Hematopoietic Stem Cells (ST-HSC). In what follows x_1 denotes the density of short-term stem-like healthy cells, x_2 the density of mature healthy cells, x_3 the density of short-term stem-like leukemic cells, x_4 the density of mature leukemic cells.

The time necessary for a ST-HSC to complete a cycle of self-renewal, asymmetric division or differentiation is τ_{1l} for leukemic cells and τ_{1h} for the healthy ones, while the time necessary for the maturation of leukocytes is denoted by τ_{2l} in the case of leukemic cells and τ_{2h} for the healthy ones.

As we mentioned in the introduction, experimental evidences attest that the imatinib therapy affects primarily the proliferation rate and secondary, the apoptosis rate. In view of this fact, we consider the treatment functions $f_u = \dfrac{1}{1-u}$ and $f_{1a} = (\gamma_{1h} - \gamma_{1l})\, u_1$, $f_{2a} = c\gamma_{2h} u_2$, with $u, u_1, u_2 : [0,T] \to [0,1]$, where $u(t), u_1(t), u_2(t)$ are the treatment effects. The action of treatment on the proliferation rate will be considered through f_u in the function of self-renew β_l and in the function of differentiation or asymmetric division k_l. Note that, in this way, both β_l and k_l became decreasing functions of u. If no drug is given (i.e. $u(t) = 0$) then $\beta_l((x_1 + y_1)f_u) = \beta_l(x_1 + y_1), k_l((x_1 + y_1)f_u) = k_l(x_2 + y_2)$ and also, a maximal effect happens for $u(t) = 1$ when the process of division essentially stops ($\beta_l \equiv k_l \equiv 0$).

Treatment will be consider to act only on the leukemic stem cells compartment. The treatment acts on the apoptosis of mature CML cells through f_{2a} and on the apoptosis of stem-like CML cells through f_{1a}, restoring this rate to a value closed to the mortality rate of healthy cells. From the law of the mass, we have $\tilde{f}_{1a} = \int_{t-\tau_{1l}}^{t} u_1(s)ds$. The optimal control model is

$$\begin{aligned}
\dot{x}_1 &= f_1(x_1, x_2, y_1, y_2, x_{1\tau_{1h}}, x_{2\tau_{1h}}, y_{1\tau_{1h}}, y_{2\tau_{1h}}) \\
\dot{x}_2 &= f_2(x_2, x_{1\tau_{2h}}, x_{2\tau_{2h}}, y_{2\tau_{2h}}) \\
\dot{y}_1 &= f_3(t, x_1, x_2, y_1, y_2, x_{1\tau_{1l}}, x_{2\tau_{1l}}, y_{1\tau_{1l}}, y_{2\tau_{1l}}, u_1, u, u_{\tau_{1l}}) \\
\dot{y}_2 &= f_4(y_2, x_{2\tau_{2l}}, y_{1\tau_{2l}}, y_{2\tau_{2l}}, u_2, u_{\tau_{2l}})
\end{aligned} \tag{1}$$

where

$$\begin{aligned}
f_1 = {} & -\gamma_{1h}x_1 - (\eta_{1h} + \eta_{2h})k_h(x_2 + y_2)x_1 - (1 - \eta_{1h} - \eta_{2h})\beta_h(x_1 + y_1)x_1 + \\
& + 2e^{-b_{1h}\tau_{1h}}(1 - \eta_{1h} - \eta_{2h})\beta_h(x_{1\tau_{1h}} + y_{1\tau_{1h}})x_{1\tau_{1h}} + \\
& + \eta_{1h}e^{-b_{1h}\tau_{1h}}k_h(x_{2\tau_{1h}} + y_{2\tau_{1h}})x_{1\tau_{1h}} \\
f_2 = {} & -\gamma_{2h}x_2 + A_h k_h(x_{2\tau_{2h}} + y_{2\tau_{2h}})x_{1\tau_{2h}} \\
f_3 = {} & -(\gamma_{1l} + \mathbf{f}_{1a})y_1 - [(\eta_{1l} + \eta_{2l})k_l((x_2 + y_2)\,\mathbf{f}_u) \\
& \qquad + (1 - \eta_{1l} - \eta_{2l})\beta_l((x_1 + y_1)\,\mathbf{f}_u)]\, y_1 + \\
& + [2e^{-b_{1l}\tau_{1l} - \tilde{\mathbf{f}}_{1a}}(1 - \eta_{1l} - \eta_{2l})\beta_l((x_{1\tau_{1l}} + y_{1\tau_{1l}})\,\mathbf{f}_{\mathbf{u}_{\tau_{1l}}}) + \\
& + \eta_{1l}e^{-b_{1l}\tau_{1l} - \tilde{\mathbf{f}}_{1a}}k_l((x_{2\tau_{1l}} + y_{2\tau_{1l}})\,\mathbf{f}_{\mathbf{u}_{\tau_{1l}}})]y_{1\tau_{1l}} \\
f_4 = {} & -(\gamma_{2l} + \mathbf{f}_{2a})\, y_2 + A_l k_l((x_{2\tau_{2l}} + y_{2\tau_{2l}})\,\mathbf{f}_{\mathbf{u}_{\tau_{2l}}})y_{1\tau_{2l}}
\end{aligned}$$

subject to minimization of the cost functional

$$\min J(u), \tag{2}$$

where

$$J(u) = ay_1(T) + by_2(T) + \int_0^T \left[u^2(t) + u_1^2(t) + u_2^2(t)\right] dt$$

with $g(y(T)) = ay_1(T) + by_2(T)$ the weighted sum of the final tumor population and $L(u(t)) = \int_0^T \left[u^2(t) + u_1^2(t) + u_2^2(t)\right] dt$, the cumulative drug toxicity.

In this paper we denote $X_\tau = X(t-\tau)$, where $X = (x_1, x_2, y_1, y_2)$. The history of the state variables is given by

$$X(\theta) = \varphi(\theta),\ \theta \in [-\tau_{\max}, 0],\ \tau_{\max} = \max(\tau_{1h}, \tau_{2h}, \tau_{1l}, \tau_{2l}).$$

It is not difficult to see that if the initial conditions have all components positive functions, the solutions of the system will have positive components on all the interval of existence. Indeed, if $x_1(\theta) > 0$ for any $\theta \in [-\tau, 0]$ and there exists $T > 0$ such that $x_1(T) = 0$ the derivative $\dot{x}_1(T)$ will be positive and this leads to a contradiction. The same argument works for the other components, too.

We assume that: a percentage $\eta_{1\alpha}$, $\alpha = h, l$ of stem-like cells population is supposed to undergo asymmetric division; a percentage $\eta_{2\alpha}$, $\alpha = h, l$ of the population differentiate symmetrically and the percentage $(1-\eta_{1\alpha}-\eta_{2\alpha})$, $\alpha = h, l$ of the population is supposed to self-renew (see also [13]).

Furthermore,it is assumed that homeostatic mechanisms maintain the hematopoietic stem cell population at a constant level. In this respect, the rate of self-renewal is given by a Hill function

$$\beta_\alpha(X) = \beta_{0\alpha} \frac{\theta_1^m}{\theta_1^m + X^m}, \alpha = h, l$$

and the rate of differentiation, through symmetric or asymmetric division is supposed to be dictated, through a feedback law, by

$$k_\alpha(X) = k_{0\alpha} \frac{\theta_2^n}{\theta_2^n + X^n}, \alpha = h, l.$$

Because in this paper we consider competition between healthy and CML cell populations, both this rates will depend on the sum of stem-like respectively mature populations (similar approaches on competition were modeled in [22,23]).

For $\alpha = h, l$, the other parameter are defined as follows: $b_{1\alpha}$ accounts for the death rate of stem cells and a positive K_α for the loss rate due to differentiation into other cell lines - the resulting loss rate is denoted as $\gamma_{1\alpha} = K_\alpha + b_{1\alpha}$; $\beta_{0\alpha}$ and $k_{0\alpha}$ represent the maximal rate of self-renewal, respectively of asymmetric division or differentiation into leukocyte line; θ_i, $i = 1, 2$, is the value for which β_α, respectively k_α attains half of their maximum value; $\gamma_{2\alpha}$ is the mortality of mature cells; A_α is an amplification factor of mature cells due to differentiation; m is the parameter controlling the sensitivity of the mitotic re-entry rate β_α to changes in the size of G_0 and n is the parameter controlling the sensitivity of the asymmetric division or differentiation rate k_α to changes in the size of mature population.

Existence of an optimal control. The existence of an optimal control results from transforming the problem into an optimal control problem for a system of ordinary differential equations (see next section) whose solutions will be bounded together with their derivatives on compact intervals (see [5]).

3 Discretization of the Optimal Control Problem

In this section, we apply the numerical procedure from Gollmann et al. [10], in order to solve the delay optimal control problem (1)+(2) (see also [8,9]). For that matter, we write the cost functional in the Mayer form

$$J(u,y) = h(y(T)),\ y = (y_1, y_2) \in R^2.$$

In our case, the reduction of the more general cost functional (2) to Mayer form, proceeds by the introduction of the additional state variable z through the delayed equation

$$\dot{z}(t) = L(u(t)) \ \text{ so } \ \dot{z}(t) = u^2(t) + u_1^2(t) + u_2^2(t), \ \ z(0) = 0.$$

Then, the cost functional (2) is rewritten as

$$J(u,y,z) = g(y(T)) + z(T).$$

In the following, let $\tau > 0$ such that $\tau_{1h} = k_1\tau$, $\tau_{2h} = k_2\tau$, $\tau_{1l} = k_3\tau$, $\tau_{2l} = k_4\tau$, $k_i \in N^*, i = \overline{1,4}$, $T = N\tau$ and use the Euler integration method with a uniform step size $\tau > 0$. Of course, τ can be refined in order to obtain an appropriate smaller step-size.

Using the grid points $t_i = i\tau$, $i = \overline{0,N}$ and the approximations $x_1(t_i) \simeq x_{1i} \in R$, $x_2(t_i) \simeq x_{2i} \in R$, $y_1(t_i) \simeq y_{1i} \in R$, $y_2(t_i) \simeq y_{2i} \in R$, $u(t_i) \simeq u_i$, $u_1(t_i) \simeq u_{1i}$ and $u_2(t_i) \simeq u_{2i}$, the treatment function f_{1a} becomes $\sum_{j=1}^{k_3} u_{1_{i-j}}\tau$ and the delay control problem (1)+(2) is transformed into the nonlinear programming problem (NLP)

$$Minimize\ J = g(x_N, y_N) + z_N \tag{3}$$

subject to

$$\begin{cases} x_{1i} - x_{1i+1} + \tau f_1(x_{1i}, x_{2i}, y_{1i}, y_{2i}, x_{1i-k_1}, x_{2i-k_1}, y_{1i-k_1}, y_{2i-k_1}) = 0 \\ x_{2i} - x_{2i+1} + \tau f_2(x_{2i}, x_{1i-k_2}, x_{2i-k_2}, y_{2i-k_2}) = 0 \\ y_{1i} - y_{1i+1} + \tau f_3(x_{1i}, x_{2i}, y_{1i}, y_{2i}, x_{1i-k_3}, x_{2i-k_3}, y_{1i-k_3}, \\ \qquad\qquad y_{2i-k_3}, u_i, u_{1i}, u_{1_{i-1}}, .., u_{1_{i-k_3}}, u_{i-k_3}) = 0 \\ y_{2i} - y_{2i+1} + \tau f_4(y_{2i}, x_{2i-k_4}, y_{1i-k_4}, y_{2i-k_4}, u_{2i}, u_{i-k_4}) = 0 \\ z_i - z_{i+1} + \tau\left(u_i^2 + u_{1i}^2 + u_{2i}^2\right) = 0 \end{cases} \tag{4}$$

$$\begin{aligned} -u_i \le 0, u_i - 1 \le 0, \\ -u_{1i} \le 0, u_{1i} - 1 \le 0, \\ -u_{2i} \le 0, u_{2i} - 1 \le 0. \end{aligned} \tag{5}$$

$$i = \overline{0, N-1}.$$

Herein, the initial value profile φ_1,φ_2, φ_3 and φ_4 gives the values

$$\begin{aligned} x_{1_{-i}} &:= \varphi_1(-i\tau),\ i = \overline{0, k_1} \\ x_{2_{-i}} &:= \varphi_2(-i\tau),\ i = \overline{0, k_2} \\ y_{1_{-j}} &:= \varphi_3(-j\tau),\ j = \overline{0, k_3} \\ y_{2_{-j}} &:= \varphi_4(-j\tau),\ j = \overline{0, k_4} \end{aligned}$$

The variable to be optimized is represented by the vector

$$\begin{aligned} w = \big(&u_0, u_{1_0}, u_{2_0}, x_{1_1}, x_{2_1}, y_{1_1}, y_{2_1}, z_1, ..., u_{N-1}, u_{1_{N-1}}, u_{2_{N-1}}, \\ &x_{1_N}, x_{2_N}, y_{1_N}, y_{2_N}, z_N\big) \in R^{8N}. \end{aligned}$$

The numerical procedure described above is applied in the next section and the graphs for states and control for different sets of parameters are obtained.

4 Numerical Results and Simulations

In the following figures, we plotted the trajectories of the healthy, respectively CML cell populations for the competition system, showing a comparison between the dynamics of a system without treatment and the dynamics of a system subject to optimal control of treatment. To solve the problem of optimal control the Matlab solver for NLP problems fmincon was used, selecting the 'interior-point' solver.

In all figures, for the healthy cell populations, we choose the same set of parameters value: $\eta_{1h} = 0.7, \eta_{2h} = 0.1, \tau_{1h} = 2, \tau_{2h} = 4, \gamma_{1h} = 0.1, \gamma_{2h} = 2.4, A_h = 922$, $\beta_{0h} = 1.77$, $k_{0h} = 0.1$, $\theta_{1h} = 0.5 \cdot 10^6$, $\theta_{2h} = 0.36 \cdot 10^8$. For leukemic cell populations, we consider alteration of the value of the following parameters:

- smaller percent of asymmetric division (η_{1l});
- bigger percent of self-renewal ($1 - \eta_{1l} - \eta_{2l}$);
- lower rate of apoptosis of leukemic stem cells (γ_{1l});
- lower rate of apoptosis of leukemic mature cells (γ_{2l});
- enhanced differentiation (A_l).

In the following example (Figs. 1 and 2) all these features were modified.

If we maintain the configuration of parameters from the previous example but consider that the percentage of self renewal of leukemic cells is the same as the percentage of self renewal of healthy cells, we see another manifestation of the disease, for which the treatment dose is lower than in the previous case (Figs. 3 and 4).

5 Discussion

The plots of optimal controls (Figs. 2 and 4) exhibit an optimal control effect almost constant at 0.5 respectively at 0.25 until the 90th day for all controls. One can intuitively expect that the treatment effect is proportional with the

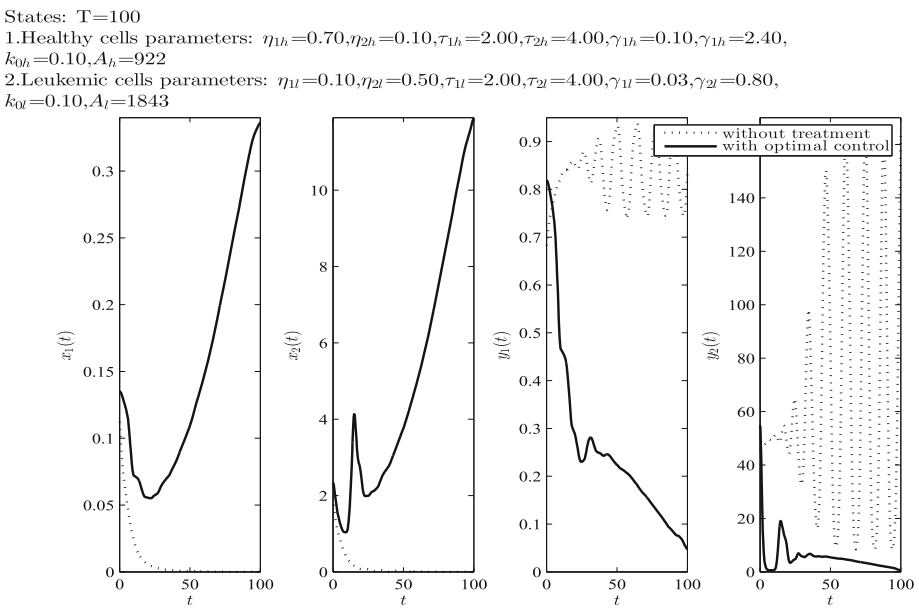


Fig. 1. Comparison between the dynamics of a system without treatment and with with optimal control of treatment. The results correspond to the best local minimal solution obtained for different initial guesses (the value of cost functional was improved from 1597 to 901)

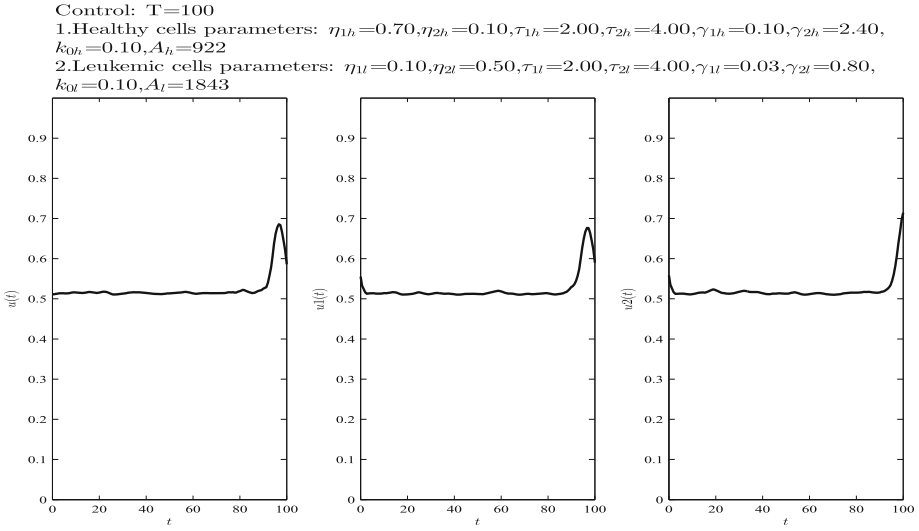


Fig. 2. Optimal control of treatment. The controls u, u1, u2 represent the influence of drug on the proliferation rate and apoptosis. One can observe that the drug influence is almost constant and the evolution of u, u1 and u2 are similar if parameters a an b of cost functional are 1.

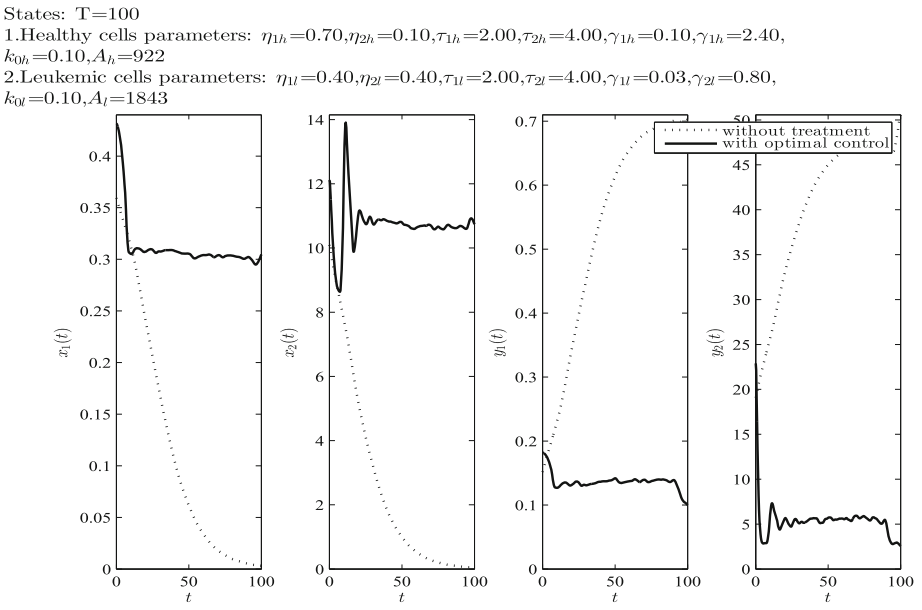


Fig. 3. Comparison between the dynamics of a system without treatment and with with optimal control of treatment. The results correspond to the best local minimal solution obtained for different initial guesses (the value of cost functional was improved from 641 to 260)

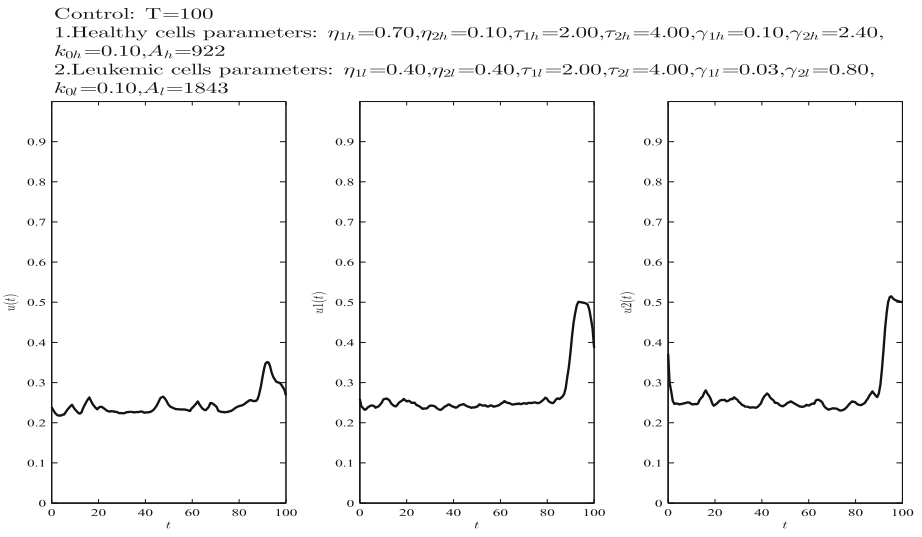


Fig. 4. Optimal control of treatment. The controls u, u1, u2 represent the influence of drug on the proliferation rate and apoptosis. One can observe that the drug influence is almost constant and the evolution of u, u1 and u2 are similar.

cell concentration of imatinib; however, in order to study the influence of the prescribed drug concentration on the population's time-evolution, imatinib pharmacokinetics (PK) and pharmacodynamics (PD) need to be taken into account (see also [20]). Nevertheless, for an optimal effect of treatment, the prescribed dose should be adapted in view of the disease parameters of a certain patient. In that respect, a competition model of healthy vs. CML cell population that takes into account the influence of PK and PD of imatinib, is subject of further research.

References

1. Adimy, M., Crauste, F., Halanay, A., Neamţu, M., Opriş, D.: Stability of limit cycles in a pluripotent stem cell dynamics model. Chaos, Solitons Fractals **27**(4), 1091–1107 (2006)
2. Adimy, M., Crauste, F., Ruan, S.: A mathematical study of the hematopoiesis process with application to chronic myelogenous leukemia. SIAM J. Appl. Math. **65**(4), 1328–1352 (2005)
3. Ainseba, B., Benosman, C.: Optimal control for resistance and suboptimal response in CML. Math. Biosci. **227**, 81–93 (2010)
4. Beckman, J., Scheitza, S., Wernet, P., Fischer, J., Giebel, B.: Asymmetric cell division within the human hematopoietic stem and progenitor cell compartment: identification of asymetrically segregating proteins. Blood **12**(109), 5494–5501 (2007)
5. Benosman, C.: Control of the dynamics of chronic myeloid leukemia by imatinib. Ph.D. thesis (2010)
6. Colijn, C., Mackey, M.C.: A mathematical model of hematopoiesis - I. Periodic chronic myelogenous leukemia. J. Theor. Biol. **237**, 117–132 (2005)
7. Deiniger, M.W.N., Goldman, J.M., Lydon, N., Melo, J.V.: The tyrosine kinase inhibitor CGP57148B selectively inhibits the growth of BCRABL positive cells. Blood **90**, 3691–3698 (1997)
8. Gollmann, L., Kern, D., Maurer, H.: Optimal control problems with control and state delays and applications to growth processes. In: IIASA Symposium on Applications of Dynamic Systems to Economic Growth with Environment, Luxemburg, 7–8 November 2008
9. Gollmann, L., Kern, D., Maurer, H.: Optimal control problems with delays in state and control variables subject to mixed control state constraints. Optim. Control Appl. Methods **30**, 341–365 (2009)
10. Gollmann, L., Maurer, H.: Theory and applications of optimal control problems with multiple time-delays. J. Ind. Manage. Optim. **10**(2), 413–441 (2014)
11. Gottschalk, S., Anderson, N., Hainz, C., et al.: Imatinib (STI571)-mediated changes in glucose metabolism in human leukemia BCR-ABL-positive cells. Clin. Cancer Res. **10**, 6661–6668 (2004)
12. Halanay, A.: Periodic solutions in mathematical models for the treatment of chronic myelogenous leukemia. Math. Model. Nat. Phenom. **7**(1), 235–244 (2012)
13. Halanay, A., Candea, D., Radulescu, I.R.: Stability analysis of equilibria in a delay differential equations model of CML including asymmetric division and treatment. Math. Comput. Simul. (2014, to apper). Elsevier
14. Kim, P., Lee, P., Levy, D.: Dynamics and potential impact of the immune response to chronic myelogenous leukemia. PLoS Comput. Biol. **4**(6), e1000095 (2008)

15. Kim, P., Lee, P., Levy, D.: A theory of immunodominance and adaptive regulation. Bull. Math. Biol. **73**(7), 1645–65 (2011)
16. Mackey, M.C., Ou, C., Pujo-Menjouet, L., Wu, J.: Periodic oscillations of blood cell population in chronic myelogenous leukemia. SIAM J. Math. Anal. **38**, 166–187 (2006)
17. Marciniak-Czochra, A., Stiehl, T., Wagner, W.: Modeling of replicative senescence in hematopoietic development. Aging **1**(8), 723–732 (2009)
18. Michor, F., Hughes, T., Iwasa, Y., Branford, S., Shah, N.P., Sawyers, C., Novak, M.: Dynamics of chronic myeloid leukemia. Nature **435**, 1267–1270 (2005)
19. Moore, H., Li, N.K.: A mathematical model for chronic myelogenous leukemia (CML) and T-cell interaction. J. Theor. Biol. **227**, 513–523 (2004)
20. Radulescu, I.R., Candea, D., Halanay, A.: Stability and bifurcation in a model for the dynamics of stem-like cells in leukemia under treatment. Am. Inst. Phys. Proc. **1493**, 758–763 (2012)
21. Reya, T.: Regulation of hematopoietic stem cell self-renewal. Recent Prog. Horm. Res. **58**, 283–295 (2003)
22. Stiehl, T., Marciniak-Czochra, A.: Mathematical modeling of leukemogenesis and cancer stem cell dynamics. Math. Model. Nat. Phenom. **7**(1), 166–202 (2012)
23. Tang, M., Foo, J., Gonen, M., Mahon, F.-X., Michor, F.: Selection pressure exerted by imatinib therapy leads to disparate outcomes of imatinib discontinuation trials. Haematologica **97**(10), 1553–1561 (2012)
24. Tomasetti, C., Levi, D.: Role of symmetric and asymmetric division of stem cells in developing drug resistance. PNAS **17**(39), 16766–16771 (2010)

More Safe Optimal Input Signals for Parameter Estimation of Linear Systems Described by ODE

Ewaryst Rafajłowicz(✉) and Wojciech Rafajłowicz

Institute of Computer Engineering Control and Robotics, Wroclaw, Poland
ewaryst.rafajlowicz@pwr.wroc.pl

Abstract. Our starting point is the ascertainment that D-optimal input signals recently considered by the same authors [12] can be too dangerous for applying them to real life system identification. The reason is that they grow too fast in time. In order to obtain more safe input signals, but still leading to a good estimation accuracy of parameter estimates, we propose a quality criterion that is a mixture of D-optimality and a penalty for too fast growth of input signals in time.

Our derivations are parallel to those in [12] up to a certain point only, since we obtain different optimality conditions in the form of an integral equation. We also briefly discuss a numerical algorithm for its solution.

Keywords: Parameter estimation · Optimal input signal · D-optimality · Optimal experiment design · Integral equations

1 Introduction

Our aim is to provide optimality conditions for optimal input signals that are D-optimal for parameter estimation in linear systems described by ordinary differential equations. In opposite to our paper [12], we put emphasis not only on parameter estimation accuracy, but also on safety of input signals. This leads to different optimality conditions than those obtained in [12]. They are derived from the variational approach and they are expressed in a convenient form of integral equations.

By the lack of space, we do not discuss the selection of input signal when a feedback is present. As it was demonstrated in [6] for systems described by ODE's – the presence of feedback can be beneficial. The result presented here can be generalized to the case with a feedback, using the results on output sensitivity to parameters for systems with feedback (see [13], but this is outside the scope of this paper.

The paper was supported by the National Council for Research of Polish Government under grant 2012/07/B/ST7/01216, internal code 350914 of Wrocław University of Technology.

An erratum to this chapter is available at 10.1007/978-3-662-45504-3_35

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 267–277, 2015.
DOI: 10.1007/978-3-662-45504-3_26

As is known, problems of selecting input signal for parameter estimation in linear systems have been considered in seventies of 20-th century. The results were summarized in [9] and two monographs [5,19]. They have been concentrated mainly on the so-called frequency domain approach. This approach leads to a beautiful theory, which runs in paralel to the optimal experiment design theory (see [1]). The main message of this approach is that optimal input signal are linear combinations of a finite number of sinusoidal waves with precisely selected frequencies. We stress that sums of sinusoids can have unexpectedly large amplitudes, which can be dangerous for an identified systems. Notice that the frequency domain approach requires an infinite observation horizon. Here we assume a finite observation horizon, which leads to quite different results.

Research in this area was stalled for nearly 20 years. Recently, since the beginning of 21-st century, the interest of researchers has rapidly grown up (see [3,4,7,8,16,17] for selected recent contributions).

Related problems of algorithms for estimating parameters in PDE's and selecting allocation of sensors are also not discussed (we refer the reader to [2,10,14,15]).

2 Problem Statement

System description. Consider a system described by ODE

$$\frac{d^r\, y(t)}{d\, t^r} + a_{r-1}\, \frac{d^{r-1}\, y(t)}{d\, t^{r-1}} + \ldots + a_0\, y(t) = a_r\, u(t), \quad t \in (0,\, T] \tag{1}$$

with zero initial conditions, where $y(t)$ is the output, $u(t)$ is the input signal.

The solution $y(t;\, \bar{a})$ of (1) depends on the vector $\bar{a} = [a_0,\, a_1, \ldots,\, a_r]^{tr}$ of unknown parameters. The impulse response (response to the Dirac Delta or Green's function) of ODE (1) will be denoted later by $g(t; \bar{a})$. Notice that $g(t; \bar{a}) = 0$, $t < 0$ and $y(t;\, \bar{a}) = \int_0^T g(t - \tau; \bar{a})\, u(\tau)\, d\tau$. Remark that it is not necessary to assume differentiability of y w.r.t. $\bar{a}$, because solutions of linear ODE's are known to be analytical functions of ODE parameters.

As is well known, differential sensitivity $y(t;\, \bar{a})$ to parameter changes, can be expressed and calculated in a number of ways (see [13] for a brief summary). For our purposes it is convenient to express it through $r \times 1$ vector of sensitivities $\bar{k}(t; \bar{a}) \stackrel{def}{=} \nabla_a g(t; \bar{a})$. Then,

$$\nabla_a y(t;\, \bar{a}) = \int_0^T \bar{k}(t - \tau; \bar{a})\, u(\tau)\, d\tau. \tag{2}$$

Observations and the estimation accuracy. Available observations have the form:

$$\Upsilon(t) = y(t;\, \bar{a}) + \varepsilon(t), \quad t \in [0,\, T],$$

where $\varepsilon(t)$ is zero mean, finite variance, uncorrelated, Gaussian white noise, more precisely, $\varepsilon(t)$ is implicitly defined by $d\, W(t) = \varepsilon(t)\, dt$, where $W(t)$ is the Wiener process.

It can be shown (see [5]), that the Fisher information matrix $\mathbf{M}_T(u)$ for estimating $\bar{a}$ has the form:

$$\mathbf{M}_T(u) = \int_0^T \nabla_a y(t;\bar{a})\,(\nabla_a y(t;\bar{a}))^{tr}\,dt,$$

where $\nabla_a y(t;\bar{a})$ depends on $u(.)$ through (2). From the linearity of (1) it follows that $\mathbf{M}_T(u)$ can be expressed as follows

$$\mathbf{M}_T(u) = \int_0^T \int_0^T H(\tau,\,\nu;\,\bar{a})\,u(\tau)\,u(\nu)\,d\tau\,d\nu,$$

where $H(\tau,\,\nu;\,\bar{a}) \stackrel{def}{=} \int_0^T \bar{k}(t-\tau;\bar{a})\,\bar{k}^{tr}(t-\nu;\bar{a})\,dt$.

From the Cramer-Rao inequality we know that for any estimator $\tilde{a}$ of $\bar{a}$ we have:

$$cov(\tilde{a}) \geq [\mathbf{M}_T(u)]^{-1}. \tag{3}$$

When observation errors are Gaussian and the minimum least squares estimator is used, then the equality in (3) is asymptotically attained. Thus, it is meaningful to minimize interpretable functions of $\mathbf{M}_T(u)$, e.g., $\min Det[\mathbf{M}_T(u)]^{-1}$ w.r.t. $u(.)$, under certain constraints on $u(.)$.

Problem formulation. Define

$$\mathcal{U}_0 = \left\{ u : \int_0^T u^2(t)\,dt \leq e_1,\ Det[\mathbf{M}_T(u)] > 0 \right\},$$

where $e_1 > 0$ is a level of available (or admissible) energy of input signals. Let $C_0(0,\,T)$ denote the space of all functions that are continuous in $[0,\,T]$.

In [12] the following problem has been considered. Find $u^* \in \mathcal{U}_0 \cap C_0(0,\,T)$ for which $\min_{u\in\mathcal{U}_0} Det\left[\mathbf{M}_T^{-1}(u)\right]$, or equivalently,

$$\max_{u\in\mathcal{U}_0} \log\left[Det\left(\mathbf{M}_T(u)\right)\right] \tag{4}$$

is attained. As it was demonstrated in [12] (see also comparisons at the end of this paper), signals that are optimal in the above sense can be even more dangerous than sums of sinusoids.

For these reasons, we consider here putting additional constraints on the system output.

Typical output constraints include:

$$\int_0^T y^2(t)\,dt \leq e_2, \quad y(t) = \int_0^T g(t-\tau,\bar{a})\,u(\tau)\,d\tau, \tag{5}$$

$$\int_0^T \dot{y}^2(t)\,dt \leq e_3, \quad \dot{y}(t) = \int_0^T g'(t-\tau,\bar{a})\,u(\tau)\,d\tau, \tag{6}$$

$$\int_0^T \ddot{y}^2(t)\,dt \leq e_4, \quad \ddot{y}(t) = \int_0^T g''(t-\tau,\bar{a})\,u(\tau)\,d\tau. \tag{7}$$

Later, we consider only (5), because results for (6) and (7) can be obtained from those for (5) by formal replacement of g by g' or g'', respectively.

Define $\mathcal{U}_1 = \left\{u : \int_0^T u^2(t)\,dt \le e_1,\ \int_0^T y^2(t)\,dt \le e_2,\ Det[\mathbf{M}_T(u)] > 0\right\}$, where $e_2 > 0$ is the admissible level of energy of output $y(t)$ signal that depends on $u(.)$ as in (5).

Now, our problem reads as follows: find $u^* \in \mathcal{U}_1 \cap C_0(0, T)$ for which

$$\max_{u\in\mathcal{U}_1} \log\left[Det\left(\mathbf{M}_T(u)\right)\right] \tag{8}$$

is attained.

In fact, in most cases only one of the constraints $\int_0^T u^2(t)\,dt \le e_1$ and $\int_0^T y^2(t)\,dt \le e_2$ is active, depending on the system properties and on values of e_1 and e_2. In practice, the following way of obtaining an input signal that is informative and simultaneously safe for the identified system can be proposed.

Upper level algorithm

1. For a given value of available input energy $e_1 > 0$ solve problem (4). Denote its solution by u_1^*, say. Set $\ell = 1$.
2. Calculate the output signal y_ℓ^*, corresponding to u_ℓ^*. If y_ℓ^* is safe for the system, then stop – u_ℓ^* is our solution. Otherwise, go to step 3.
3. Calculate $e_{trial} = \int_0^T [y_\ell^*(t)]^2\,dt$ and select e_2 less then e_{trial}, e.g., $e_2 = \theta\, e_{trial}$, where $0 < \theta < 1$.
4. Set $\ell = \ell + 1$ and solve problem (8). Denote its solution by u_ℓ^* and go to step 2. Notice that now the constraint for the output energy is active and the one on input energy is almost surely not active (see the discussion in the next section).

The curse of the lack of a priori knowledge. As in optimum experiment design for nonlinear (in parameters) regression estimation (see [1,11]), also here the optimal u^* depends on unknown $\bar{a}$. Furthermore, condition $\int_0^T [y(t)]^2\,dt \le e_2$ also contains unknown $\bar{a}$. The ways of circumventing these difficulties have been discussed for a long time. The following ways are usually recommended (see [11]):

1. use "the nominal" parameter values for $\bar{a}$, e.g., values from previous experiments,
2. the "worst case" analysis, i.e. solve the following problem

$$\max_{u\in\mathcal{U}_1} \min_{\bar{a}} \log\left[Det\left(\mathbf{M}_T(u;\bar{a})\right)\right],$$

   where $\mathbf{M}_T(u;\bar{a})$ is the Fisher information matrix, in which dependence on $\bar{a}$ is displayed,
3. the Bayesian approach: use prior distribution imposed on $\bar{a}$ and average $\mathbf{M}_T(u;\bar{a})$ with respect to it,
4. apply the adaptive approach of subsequent estimation and planning stages.

Later we use the "nominal" parameter values $\bar{a}$. Thus, we obtain locally optimal input signals, which are optimal in a vicinity of nominal parameters. We underline that the results are relevant also to the Bayesian and adaptive approaches in the sense that it is easy to obtain optimality conditions in a way that is similar to the one presented bellow.

3 Optimality Conditions

Define the Lagrange function

$$L(u,\gamma) = \log\left[Det\left(\mathbf{M}_T(u)\right)\right] - \gamma_1 \left(\int_0^T u^2(t)\,dt - e_1\right) - \gamma_2 \left(\int_0^T y^2(t)\,dt - e_2\right),$$

where γ_1, γ_2 are the Lagrange multipliers and $\gamma = [\gamma_1, \gamma_2]^{tr}$.

Let $u^* \in \mathcal{U}_1 \cap C_0(0, T)$ be a solution of (8) problem and let $u_\epsilon(t) = u^*(t) + \epsilon\, f(t)$, where $f \in C_0(0, T)$ is arbitrary. Then, for the Gateaux differential of L we obtain

$$\left.\frac{\partial L(u_\epsilon, \gamma)}{\partial\, \epsilon}\right|_{\epsilon=0} = 2 \int_0^T f(\nu) \left[\int_0^T \widehat{ker}(\tau, \nu, u^*)\, u^*(\tau)\, d\tau - \gamma_1\, u^*(\nu)\right] d\nu,$$

where, for $u \in \mathcal{U}_0$, we define:

(a) $\widehat{ker}(\tau, \nu, u) \stackrel{def}{=} ker(\tau, \nu, u) - \gamma_2\, G(\tau, \nu)$,
(b) $G(\tau, \nu) \stackrel{def}{=} \int_0^T g(t-\tau, \bar{a})\, g(t-\nu, \bar{a})\, dt$,
(c) $ker(\tau, \nu, u) \stackrel{def}{=} trace\left[\mathbf{M}_T^{-1}(u)\, H(\tau, \nu, \bar{a})\right]$.

The symmetry of kernel $\widehat{ker}(\tau, \nu, u)$ was used in calculating $\left.\frac{\partial L(u_\epsilon, \gamma)}{\partial\, \epsilon}\right|_{\epsilon=0}$. Notice that $\widehat{ker}(\tau, \nu, u)$ depends also on γ_2 but this is not displayed in the notation.

If u^* is optimal, then for each $f \in C_0(0, T)$ we have $\left.\frac{\partial L(u_\epsilon, \gamma)}{\partial\, \epsilon}\right|_{\epsilon=0} = 0$. Then, by the fundamental lemma of the calculus of variation we obtain.

Proposition 1. *If u^* solves problem (8), then it fulfils the following integral equation*

$$\int_0^T \left[ker(\tau, \nu, u^*) - \gamma_2\, G(\tau, \nu)\right] u^*(\tau)\, d\tau = \gamma_1\, u^*(\nu), \quad \nu \in (0, T). \tag{9}$$

Notice that the constraints:

– input energy constraint $\int_0^T u^2(t)\, dt \le e_1$ and
– output energy constraint $\int_0^T y^2(t)\, dt \le e_2$

can be simultaneously active at the optimal solution u^* in a very special case. Namely, if

$$\int_0^T [u^*(t)]^2\, dt = e_1, \tag{10}$$

then we must also have

$$\int_0^T \int_0^T G(\tau,\, \nu)\, u^*(\tau)\, u^*(\nu)\, d\tau\, d\nu = e_2. \tag{11}$$

In other words, simultaneous equality of both constraints (10) and (11) is not a generic case and without loosing generality, we can consider only the cases when only one of them[1] is active.

Case I – only input energy constraint active. This case is exactly the one considered in our paper [12]. We provide a brief summary of the results and their extension. Notice that in this case $\gamma_2 = 0$ and we consider problem (4). For simplicity of formulas we set $e_1 = 1$.

Proposition 2. *If u^* solves problem (4), then it fulfils the following integral equation*

$$\int_0^T ker(\tau,\, \nu,\, u^*)\, u^*(\tau)\, d\tau = \gamma_1\, u^*(\nu), \quad \nu \in (0,\, T). \tag{12}$$

Furthermore, $\gamma_1 = r + 1 = dim(\bar{a})$ and this is the largest eigenvalue of the following linear eigenvalue problem[2]

$$\int_0^T ker(\tau,\, \nu,\, u^*)\, \phi(\tau)\, d\tau = \lambda\, \phi(\nu), \quad \nu \in (0,\, T). \tag{13}$$

Thus, the eigenfunction corresponding to the largest eigenvalue is a natural candidate for being the optimal input signal. In the case of multiple eigenvalues, one can consider all linear combinations of the eigenfunctions that correspond to the largest eigenvalue.

Proposition 3. *Condition (12) with $\gamma_1 = r + 1$ is sufficient for the optimality of u^* in problem (4).*

This result was announced in [12] under additional condition on T. It occurs that this condition can be removed. The proof is given in the Appendix.

Algorithm – Case I. As one can notice, (13) is nonlinear, because of the dependence of $ker(\tau,\, \nu,\, u^*)$. The simplest algorithm that allows to circumvent this difficulty is the following:

[1] We exclude also the case that both of them are inactive, because if $u(t)$ is replaced by $\rho\, u(t)$ with $\rho > 1$, then $det[\mathbf{M}_T(\rho\, u)] > det[\mathbf{M}_T(\, u)]$.

[2] As is known [18], the eigenvalue problem for linear integral equation with nonnegative definite and symmetric kernel has the following solution: its eigenvalues are real and nonnegative, while the corresponding eigenfunctions are orthonormal in $L_2(0,\, T)$ (or can be orthonormalized, if there are multiple eigenvalues).

Start with arbitrary $u_0 \in \mathcal{U}_0$, $||u_0|| = 1$ and iterate for $p = 0, 1, \ldots$

$$\hat{u}_{p+1}(\nu) = \frac{1}{r+1} \int_0^T ker(\tau, \nu, u_p)\, u_p(\tau)\, d\tau, \quad u_{p+1} = \hat{u}_{p+1}/||\hat{u}_{p+1}||$$

until $||u_{p+1} - u_p|| < \delta$, where $\delta > 0$ is a preselected accuracy.

The proof of convergence can be based on the fixed point theorem, but this is outside the scope of this paper.

This algorithm has been used for solving examples in the next section. It was convergent in several or at most several dozens of steps. The Nyström method with the grid step size 0.01 was used to approximate integrals.

Case II – only output energy constraint active. If only the output energy constraint active then $\gamma_1 = 0$ and Proposition 1 immediately implies.

Proposition 4. *If u^* solves problem (4) with the input energy constraint inactive, then it fulfils the following integral equation for $\nu \in (0, T)$*

$$\int_0^T ker(\tau, \nu, u^*)\, u^*(\tau)\, d\tau = \gamma_2 \int_0^T G(\tau, \nu)\, u^*(\tau)\, d\tau. \tag{14}$$

Furthermore, $\gamma_2 = (r+1)/e_2$.

The last fact follows from the multiplication of the both sides of (14) by $u^*(\nu)$ and their subsequent integration. One should also observe that

$$\int_0^T \int_0^T G(\tau, \nu)\, u^*(\tau)\, u^*(\nu)\, d\tau\, d\nu = \int_0^T [y^*(t)]^2\, dt = e_2.$$

For solving (14) one can use iterations analogous to the Algorithm - Case I.

The discussion of sufficiency of (14) is outside the scope of this paper. It is however worth noticing that the following linear, generalized eigenvalue problem is associated with (14): find non-vanishing eigenfunctions and eigenvalues of the following equation:

$$\int_0^T ker(\tau, \nu, u^*)\, \psi(\tau)\, d\tau = \gamma_2 \int_0^T G(\tau, \nu)\, \psi(\tau)\, d\tau, \quad \nu \in (0, T). \tag{15}$$

4 Example

Consider the system $\ddot{y}(t) + 2\,\xi\,\dot{y}(t) + \omega_0^2\, y(t) = \omega_0\, u(t)$ with known resonance frequency ω_0 and unknown damping parameter ξ to be estimated, $\dot{y}(0) = 0$, $y(0) = 0$. Its impulse response has the form: $g(t;\xi) = \exp(-\xi\, t)\, \sin(\omega_0\, t)$, $t > 0$, while its sensitivity $k(t;\xi) = \frac{d\, g(t;\xi)}{d\, \xi}$ has the form: $k(t;\xi) = -t\, \exp(-\xi\, t)\, \sin(\omega_0\, t)$, $t > 0$.

Our starting point for searching informative but safe input signal for estimating ξ is $u_0(t) = 0.14\, sin(3\, t)$, $t \in [0,\, 2.5]$. It provides $\mathbf{M}_T = 0.013$ and $\int y^2 = 25.8$ at $\xi_0 = 0.1$, which is our nominal value. Firstly, problem (4) has been solved.

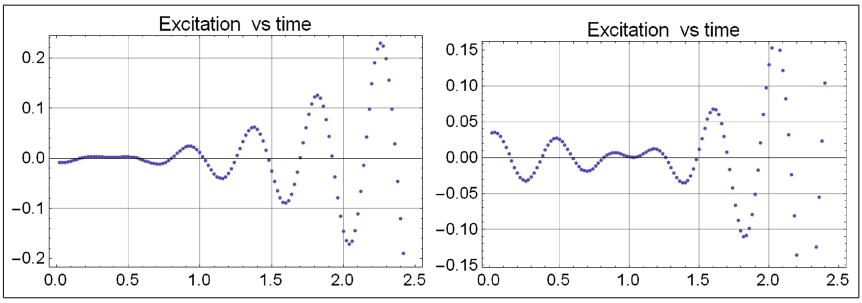


Fig. 1. Left panel – more aggressive input signal (no output constraints, right panel – more safe input signal (with output power constraint) obtained for a system described in Sect. 4.

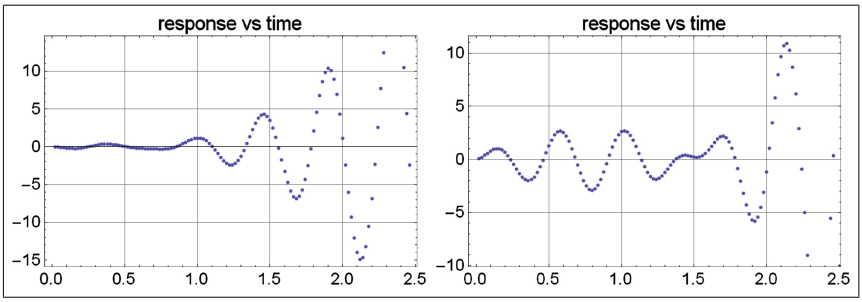


Fig. 2. Left panel – more aggressive output signal (no output constraints, right panel – more safe output signal (with output power constraint) obtained for a system described in Sect. 4.

The results are shown at the left panels of Figs. 1 and 2 for input signal and the corresponding output with $\int y^2 = 68.7$, respectively. This input signal is much more informative than u_0 – it provides $\mathbf{M}_T = 0.745$, but is too aggressive – its largest amplitude is about 0.25.

Then, according to the proposed methodology, the constraint $\int y^2 \leq 53.2$ has been added and problem (8) has been solved. The results are shown at the right panels of Figs. 1 and 2 for input signal and the corresponding output with $\int y^2 = 53.2$, respectively. This input signal is less aggressive – its maximum is 0.15 – and only slightly less informative $\mathbf{M}_T = 0.715$ than more aggressive input signal described above. The relative information efficiency of these two signals is 96%. Thus, we have reduced the largest amplitude by 60%, the output energy by 77%, while D-efficiency dropped by 4%. It is worth to mention that the more safe input signal provides 56 times better information content than our initial guess u_0.

Appendix

Proof of Proposition 3. Notice that $\gamma_2 = 0$ in this case. By direct calculations for $\frac{\partial L(u_\epsilon,\gamma)}{\partial\,\epsilon}$ we have the following expression:

$$\frac{\partial L(u_\epsilon,\gamma)}{\partial\,\epsilon} = 2\int_0^T f(\nu)\left[\int_0^T ker(\tau,\nu,u_\epsilon)\,(u^*(\tau)+\epsilon\,f(\tau))\,d\tau - \gamma\,u_\epsilon(\nu)\right]d\nu$$

Before obtaining $\frac{\partial^2 L(u_\epsilon,\gamma)}{\partial\,\epsilon^2}$ it is worth considering $K_\epsilon(\tau,\,\nu,\,u_\epsilon) \stackrel{def}{=} \left.\frac{\partial ker(\tau,\nu,u_\epsilon)}{\partial\,\epsilon}\right|_{\epsilon=0}$

$$K_\epsilon(\tau,\,\nu,\,u_\epsilon) = \int_0^T \bar{k}^{tr}(t'-\tau;\bar{a})\,\frac{\partial\,M_T^{-1}(u_\epsilon)}{\partial\,\epsilon}\Big|_{\epsilon=0}\bar{k}(t'-\nu;\bar{a})\,dt' \tag{16}$$

Using the well known formula $\frac{\partial\,B^{-1}(\epsilon)}{\partial\,\epsilon} = -B^{-1}(\epsilon)\,\frac{\partial\,B(\epsilon)}{\partial\,\epsilon}\,B^{-1}(\epsilon)$, valid for differentiable matrix valued functions $B(\epsilon)$ that are nonsingular and symmetric, we obtain

$$\frac{\partial\,M_T^{-1}(u_\epsilon)}{\partial\,\epsilon}\Big|_{\epsilon=0} = -\,M_T^{-1}(u^*)\,\mathcal{G}(u^*,\,f)\,M_T^{-1}(u^*) \tag{17}$$

where $\mathcal{G}(u^*,\,f)$ is an $R\times R$ symmetric matrix defined as follows

$$\mathcal{G}(u^*,\,f) \stackrel{def}{=} \int_0^T\left[\int_0^T \bar{k}(t-\nu';\bar{a})\,u^*(\nu')\,d\nu'\int_0^T \bar{k}^{tr}(t-\tau';\bar{a})\,f(\tau')\,d\tau' + \right. \tag{18}$$

$$\left.\int_0^T \bar{k}(t-\nu;\bar{a})\,f(\nu)\,d\nu\int_0^T \bar{k}^{tr}(t-\tau;\bar{a})\,u^*(\tau)\,d\tau\right]dt$$

Define $\bar{f} = \int_0^T f(\nu)\,d\nu$, $\overline{f^2} = \int_0^T f^2(\nu)\,d\nu$. Differentiation of $L(u_\epsilon,\,\gamma_1)$ w.r.t. ϵ yields $\left.\frac{\partial^2 L(u_\epsilon,\gamma_1)}{\partial\,\epsilon^2}\right|_{\epsilon=0} =$

$$= 2\left(\bar{f}^2 - \gamma_1\,\overline{f^2}\right) + 2\int_0^T\int_0^T K_\epsilon(\tau,\,\nu,\,u^*)\,f(\nu)\,u^*(\tau)\,d\tau\,d\nu \tag{19}$$

Subsequent substitutions of (18) into (17) and then to (16) plus tedious calculations lead to the following expression for the second summand in (19)

$$-\,2\,trace\left[M_T^{-1}(u^*)\,Z\,\left(Z\;M_T^{-1}(u^*)\right)^{tr}\right], \tag{20}$$

where $Z \stackrel{def}{=} \int_0^T Y(t)\,F^{tr}(t)\,dt$, while $Y(t) \stackrel{def}{=} \int_0^T \bar{k}(t-\tau)\,u^*(\tau)\,d\tau$, $F(t) \stackrel{def}{=} \int_0^T \bar{k}(t-\nu)\,f(\nu)\,d\nu$. The matrix in the square brackets in (20) is nonnegative definite. Thus, the whole expression is negative or zero.

From the obvious inequality $\int_0^T\left(\sqrt{\gamma_1}\,f(t)-\bar{f}\right)^2 dt \geq 0$ we immediately obtain $\left(\bar{f}^2 - \gamma_1\,\overline{f^2}\right) \leq (1-T-2\sqrt{\gamma_1})\,\bar{f}^2$ If $(1-2\sqrt{\gamma_1}-T) \leq 0$, then the

expression $\left(\overline{f}^2 - \gamma_1 \overline{f^2}\right)$ is nonpositive for all admissible non-constant functions $f \neq 0$, which yields that (19) is negative, proving sufficiency. It remains to be sure that $(1 - 2\sqrt{\gamma_1} - T) \leq 0$ or equivalently that

$$T \geq 1 - 2\sqrt{\gamma_1}. \tag{21}$$

Notice that $\gamma_1 = r + 1$. Thus, $1 - 2\sqrt{(r+1)}$ is always negative, even if only one parameter is estimated. Hence, condition (21) always holds for $T > 0$.

References

1. Atkinson, A.C., Donev, A.N., Tobias, R.D.: Optimum Experimental Designs, with SAS. Oxford University Press Inc., New York (2007)
2. Banks, H.T., Kunisch, K.: Estimation Techniques for Distributed Parameter Systems. Birkhauser, Boston (1989)
3. Gevers, M., et al.: Optimal experiment design for open and closed-loop system identification. Commun. Inf. Syst. **11**(3), 197–224 (2011)
4. Gevers, M.: Identification for control: from the early achievements to the revival of experiment design. Eur. J. Control **11**, 1–18 (2005)
5. Goodwin, G.C., Payne, R.L.: Dynamic System Identification. Experiment Design and Data Analysis, Mathematics in Science and Engineering. Academic Press, New York (1977)
6. Hjalmarsson, H., Gevers, M., De Bruyne, F.: For model-based control design, closed-loop identification gives better performance. Automatica **32**, 1659–1673 (1996)
7. Hjalmarsson, H.: System identification of complex and structured systems. Eur. J. Control **15**, 275310 (2009)
8. Hjalmarsson, H., Jansson, H.: Closed loop experiment design for linear time invariant dynamical systems via LMIs. Automatica **44**, 623636 (2008)
9. Mehra, R.K.: Optimal input signals for parameter estimation in dynamic systems, survey and new results. IEEE Trans. Automat. Control **AC–21**, 55–64 (1976)
10. Patan, M.: Optimal Sensor Network Scheduling in Identification of Distributed Parameter Systems. Lecture Notes in Control and Information Sciences, vol. 425. Springer, Heidelberg (2012)
11. Pronzato, L., Pazman, A.: Design of Experiments in Nonlinear Models. Lecture Notes in Statistics, vol. 212. Springer, Heidelberg (2013)
12. Rafajlowicz, E., Rafajlowicz, W.: A variational approach to optimal input signals for parameter estimation in systems with spatio-temporal dynamics. In: Uciski, D., Atkinson, A.C., Patan, M. (eds.) 2013 Proceedings of the 10th International Workshop in Model-Oriented Design and Analysis, Lagow, Poland. Contribution to Statistics, pp. 219–227. Springer, Heidelberg (2013)
13. Rafajlowicz, E., Rafajlowicz, W.: Control of linear extended nD sytems with minimized sensitivity to parameter uncertainties. Multidimension. Syst. Sig. Process. **24**, 637–656 (2013)
14. Skubalska-Rafajłowicz, E., Rafajłowicz, E.: Sampling multidimensional signals by a new class of quasi-random sequences, Multidimension. Syst. Sign. Process. published on-line, June 2010. doi:10.1007/s11045-010-0120-5
15. Ucinski, D.: Optimal Measurement Methods for Distributed Parameter System Identification. CRC Press, London (2005)

16. Valenzuela, P., Rojas, C., Hjalmarsson, H.: Optimal input design for non-linear dynamic systems: a graph theory approach (2013). arXiv preprint arXiv:1310.4706
17. Wahlberg, B., Hjalmarsson, H., Stoica, P.: On the performance of optimal input signals for frequency response estimation. IEEE Trans. Automa. Control **57**(3), 766–771 (2012)
18. Yosida, K.: Functional Analysis. Springer, Heidelberg (1981)
19. Zarrop, M.B.: Optimal Experimental Design for Dynamic System Identification. Lecture Notes in Control and Information Science, vol. 21. Springer, Heildelberg (1979)

Exponential Stability of Compactly Coupled Wave Equations with Delay Terms in the Boundary Feedbacks

Salah-Eddine Rebiai(✉) and Fatima Zohra Sidi Ali

LTM, Department of Mathematics, Faculty of Sciences,
University of Batna, 05000 Batna, Algeria
rebiai@hotmail.com

Abstract. We consider a linear system of compactly coupled wave equations with Neumann feedback controllers that contain delay terms. 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. This result is obtained by introducing a suitable energy function and by using an observability estimate.

Keywords: Coupled wave equations · Time delays · Boundary stabilization

1 Introduction

In [1,2], Datko et al. presented examples of infinite-dimensional second-order systems that become unstable when arbitrary small time delays occur in the damping.

Xu et al. established in [9] sufficient conditions that guarantee the exponential stability of the one-dimensional wave equation with a delay term in the boundary feedback. Nicaise and Pignotti [6] 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 [7] for the Schrödinger equation.

Motivated by the references [3,5,6,9], we investigate in this paper the problem of exponential stability for a linear system of compactly coupled wave equations with delay terms in the boundary feedbacks.

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 $\overline{\Gamma_1} \cap \overline{\Gamma_2} = \emptyset$. Furthermore, 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

$$\int_\Omega J(x)\zeta(x).\zeta(x)d\Omega \geq c \int_\Omega |\zeta(x)|^2 \, d\Omega,$$

for some constant $c > 0$ and for all $\zeta \in L^2(\Omega; \mathbb{R}^n)$,

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 278–284, 2014.
DOI: 10.1007/978-3-662-45504-3_27

$(H.2)$ $h(x).\nu(x) \leq 0$ on Γ_1,
where ν is the unit normal on Γ pointing towards the exterior of Ω.

Consider the following coupled system of two wave equations with delay terms in the boundary conditions:

$$\frac{\partial^2 u(x,t)}{\partial t^2} - \Delta u(x,t) + l(u(x,t) - v(x,t)) = 0 \qquad \text{in } \Omega \times (0,+\infty), \tag{1}$$

$$\frac{\partial^2 v(x,t)}{\partial t^2} - \Delta v(x,t) + l(v(x,t) - u(x,t)) = 0 \qquad \text{in } \Omega \times (0,+\infty), \tag{2}$$

$$u(x,0) = u_0(x), \frac{\partial u(x,0)}{\partial t} = u_1(x) \qquad \text{in } \Omega, \tag{3}$$

$$v(x,0) = v_0(x), \frac{\partial v(x,0)}{\partial t} = v_1(x) \qquad \text{in } \Omega, \tag{4}$$

$$u(x,t) = v(x,t) = 0 \qquad \text{on } \Gamma_1 \times (0,+\infty), \tag{5}$$

$$\frac{\partial u(x,t)}{\partial \nu} = -\alpha_1 \frac{\partial u(x,t)}{\partial t} - \alpha_2 \frac{\partial u(x,t-\tau)}{\partial t} \qquad \text{on } \Gamma_2 \times (0,+\infty), \tag{6}$$

$$\frac{\partial v(x,t)}{\partial \nu} = -\beta_1 \frac{\partial v(x,t)}{\partial t} - \beta_2 \frac{\partial v(x,t-\tau)}{\partial t} \qquad \text{on } \Gamma_2 \times (0,+\infty), \tag{7}$$

$$\frac{\partial u(x,t-\tau)}{\partial t} = g(x,t-\tau) \qquad \text{on } \Gamma_2 \times (0,\tau), \tag{8}$$

$$\frac{\partial v(x,t-\tau)}{\partial t} = h(x,t-\tau) \qquad \text{on } \Gamma_2 \times (0,\tau). \tag{9}$$

Physically, u and v may represent the displacements of two vibratings objects measured from their equilibrium positions, the coupling terms $\pm l(u-v)$ are the distributed springs linking the two vibrating objects. $l, \alpha_1, \alpha_2, \beta_1, \beta_2$ are positive constants, τ is the time delay, u_0, u_1, v_0, v_1, g and h are the initial data.

It is well known that in the absence of delay (*i.e.* $\alpha_2 = \beta_2 = 0$), the solution of (1)–(9) with α_1 and β_1 positive, decays exponentially to zero in the energy space $H^1_{\Gamma_1}(\Omega) \times L^2(\Omega) \times H^1_{\Gamma_1}(\Omega) \times L^2(\Omega)$ (see [5] and [3]).

The purpose of this paper is to investigate the uniform exponential stability of system (1)–(9) in the case where all the boundary damping coefficients $\alpha_1, \alpha_2, \beta_1$ and β_2 are positive. To this end, assume as in [6] that

$$\alpha_1 > \alpha_2, \beta_1 > \beta_2 \tag{10}$$

and define the energy of a solution of (1)–(9) by

$$E(t) = \frac{1}{2} \int_\Omega [|\nabla u(x,t)|^2 + \left|\frac{\partial u(x,t)}{\partial t}\right|^2 + |\nabla v(x,t)|^2 + \left|\frac{\partial v(x,t)}{\partial t}\right|^2 + l\,|u(x,t) - v(x,t)|^2]dx + \frac{1}{2} \int_{\Gamma_2} \int_0^1 [\mu \left|\frac{\partial u(x,t-\tau\rho)}{\partial t}\right|^2 + \xi \left|\frac{\partial v(x,t-\tau\rho)}{\partial t}\right|^2]d\rho\, d\Gamma \tag{11}$$

where

$$\tau\alpha_2 < \mu < \tau(2\alpha_1 - \alpha_2) \tag{12}$$

and

$$\tau\beta_2 < \xi < \tau(2\beta_1 - \beta_2) \tag{13}$$

We show that if $\{\Omega, \Gamma_1, \Gamma_2\}$ satisfies $(H.1)$ and $(H.2)$, then there is an exponential decay rate for $E(t)$.The proof of this result is based on Carleman estimates for a system of coupled nonconservative hyperbolic systems established by Lasieka and Triggiani in [4] and on compactness-uniqueness arguments.

The main result of this paper can be stated as follows.

Theorem 1. *Assume* $(H1)$, $(H.2)$, (10),(12) *and* (13). *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 Sect. 3. In Sect. 2, we study the well-posedness of system (1)–(9) using semigroup theory.

2 Well-Posedness of System (1)–(9)

Inspired from [6] and [7], we introduce the auxilliary variables

$$y(x,\rho,t) = \frac{\partial u(x,t-\tau\rho)}{\partial t}$$
$$z(x,\rho,t) = \frac{\partial v(x,t-\tau\rho)}{\partial t}$$

With these new unknowns, system (1)–(9) is equivalent to

$$\frac{\partial^2 u(x,t)}{\partial t^2} - \Delta u(x,t) + l(u(x,t) - v(x,t)) = 0 \qquad \text{in } \Omega \times (0,+\infty), \tag{14}$$

$$\frac{\partial y(x,\rho,t)}{\partial t} + \frac{1}{\tau}\frac{\partial y(x,\rho,t)}{\partial \rho} = 0 \qquad \text{on } \Gamma_2 \times (0,1) \times (0,+\infty), \tag{15}$$

$$\frac{\partial^2 v(x,t)}{\partial t^2} - \Delta v(x,t) + l(v(x,t) - u(x,t)) = 0 \qquad \text{in } \Omega \times (0,+\infty), \tag{16}$$

$$\frac{\partial z(x,\rho,t)}{\partial t} + \frac{1}{\tau}\frac{\partial z(x,\rho,t)}{\partial \rho} = 0 \qquad \text{on } \Gamma_2 \times (0,1) \times (0,+\infty), \tag{17}$$

$$u(x,t) = v(x,t) = 0 \qquad \text{on } \Gamma_1 \times (0,+\infty), \tag{18}$$

$$\frac{\partial u(x,t)}{\partial \nu} = -\alpha_1 \frac{\partial u(x,t)}{\partial t} - \alpha_2 y(x,1,t) \qquad \text{on } \Gamma_2 \times (0,+\infty), \tag{19}$$

$$\frac{\partial v(x,t)}{\partial \nu} = -\beta_1 \frac{\partial u(x,t)}{\partial t} - \beta_2 z(x,1,t) \qquad \text{on } \Gamma_2 \times (0,+\infty), \tag{20}$$

$$y(x,0,t) = \frac{\partial u(x,t)}{\partial t}, z(x,0,t) = \frac{\partial v(x,t)}{\partial t} \qquad \text{on } \Gamma_2 \times (0,+\infty), \tag{21}$$

$$u(x,0) = u_0(x), \frac{\partial u(x,0)}{\partial t} = u_1(x) \qquad \text{in } \Omega, \tag{22}$$

$$v(x,0) = v_0(x), \frac{\partial v(x,0)}{\partial t} = v_1(x) \qquad \text{in } \Omega, \tag{23}$$

$$y(x,\rho,0) = g(x,-\tau\rho),\ z(x,\rho,0) = h(x,-\tau\rho) \qquad \text{on } \Gamma_2 \times (0,1). \tag{24}$$

Denote by $\mathcal{H}$ the Hilbert space

$$\mathcal{H} = H^1_{\Gamma_1}(\Omega) \times L^2(\Omega) \times L^2(\Gamma_2; L^2(0,1)) \times H^1_{\Gamma_1}(\Omega) \times L^2(\Omega) \times L^2(\Gamma_2; L^2(0,1))$$

where

$$H^1_{\Gamma_1}(\Omega) = \{u \in H^1(\Omega) : u = 0 \text{ on } \Gamma_1\}$$

We equip $\mathcal{H}$ with the inner product

$$\left\langle \begin{pmatrix} \zeta \\ \eta \\ \theta \\ \phi \\ \chi \\ \psi \end{pmatrix}; \begin{pmatrix} \widetilde{\zeta} \\ \widetilde{\eta} \\ \widetilde{\theta} \\ \widetilde{\phi} \\ \widetilde{\chi} \\ \widetilde{\psi} \end{pmatrix} \right\rangle = \int_\Omega (\nabla\zeta(x).\nabla\widetilde{\zeta}(x) + \eta(x)\widetilde{\eta}(x))\, dx +$$

$$\mu \int_{\Gamma_2}\int_0^1 \theta(x,\rho)\widetilde{\theta}(x,\rho) d\rho\, d\Gamma + \int_\Omega (\nabla\phi(x).\nabla\widetilde{\phi}(x) + \chi(x)\widetilde{\chi}(x))\, dx +$$

$$\xi \int_{\Gamma_2}\int_0^1 \psi(x,\rho)\widetilde{\psi}(x,\rho) d\rho\, d\Gamma + l\int_\Omega (\zeta(x) - \phi(x))(\widetilde{\zeta}(x) - \widetilde{\phi}(x)) dx$$

Define in $\mathcal{H}$ a linear operator $\mathcal{A}$ by

$$D(\mathcal{A}) = \{(\zeta,\eta,\theta,\phi,\chi,\psi)^T \in H^2(\Omega) \times H^1_{\Gamma_1}(\Omega) \times L^2(\Gamma_2; H^1(0,1)) \times$$
$$H^2(\Omega) \times H^1_{\Gamma_1}(\Omega) \times L^2(\Gamma_2; H^1(0,1)); \frac{\partial\zeta}{\partial\nu} = -\alpha_1\eta - \alpha_2\theta(.,1),$$
$$\eta = \theta(.,0) \text{ on } \Gamma_2;\ \frac{\partial\phi}{\partial\nu} = -\beta_1\chi - \beta_2\psi(.,1), \chi = \psi(.,0) \text{ on } \Gamma_2\} \tag{25}$$

$$A(\zeta,\eta,\theta,\phi,\chi,\psi)^T = (\eta, \Delta\zeta + l\phi - l\zeta, -\tau^{-1}\frac{\partial\theta}{\partial\rho}, \chi, \Delta\phi - l\phi + l\zeta, -\tau^{-1}\frac{\partial\psi}{\partial\rho})^T \tag{26}$$

Then we can rewrite (14)–(24) as an abstract Cauchy problem in $\mathcal{H}$

$$\begin{cases} \frac{d}{dt}W(t) = \mathcal{A}W(t) \\ W(0) = W_0 \end{cases} \tag{27}$$

where

$$W(t) = (u(x,t), \frac{\partial u(x,t)}{\partial t}, y(x,\rho,t), v(x,t), \frac{\partial v(x,t)}{\partial t}, z(x,\rho,t))^T,$$
$$\text{and } W_0 = (u_0, u_1, g(.,-.\tau), v_0, v_1, h(.,-.\tau))^T$$

We verify that $\mathcal{A}$ is dissipative and that $\lambda I - \mathcal{A}$ is onto for a fixed $\lambda > 0$. Thus, by the Lumer-Phillips Theorem (see for instance [8]) $\mathcal{A}$ generates a strongly continuous semigroup on $\mathcal{H}$ and consequently we have

Proposition 1. *For every $W_0 \in \mathcal{H}$, problem (27) has a unique solution W whose regularity depends on the initial datum W_0 as follows:*

$$W(.) \in C([0,+\infty);\mathcal{H}) \text{ if } W_0 \in \mathcal{H},$$
$$W(.) \in C^1([0,+\infty);\mathcal{H}) \cap C([0,+\infty);D(\mathcal{A})) \text{ if } W_0 \in D(\mathcal{A}).$$

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.

Differentiating $E(t)$ with respect to time, we obtain

$$\frac{d}{dt}E(t) \leq -k\int_{\Gamma_2}\{\left|\frac{\partial u(x,t)}{\partial t}\right|^2+\left|\frac{\partial u(x,t-\tau)}{\partial t}\right|^2+\left|\frac{\partial v(x,t)}{\partial t}\right|^2+\left|\frac{\partial v(x,t-\tau)}{\partial t}\right|^2\}d\Gamma \tag{28}$$

where

$$k = \min\{\alpha_1 - \frac{\alpha_2}{2} - \frac{\mu}{2\tau}, \frac{\mu}{2\tau} - \frac{\alpha_2}{2}, \beta_1 - \frac{\beta_2}{2} - \frac{\xi}{2\tau}, \frac{\xi}{2\tau} - \frac{\beta_2}{2}\}$$

Step 2.

We rewrite

$$E(t) = \mathcal{E}(t) + E_d(t)$$

where

$$\mathcal{E}(t) = \frac{1}{2}\int_\Omega\{|\nabla u(x,t)|^2+\left|\frac{\partial u(x,t)}{\partial t}\right|^2+|\nabla v(x,t)|^2+\left|\frac{\partial v(x,t)}{\partial t}\right|^2+l\,|u(x,t)-v(x,t)|^2\}dx$$

and

$$E_d(t) = \frac{1}{2}\int_{\Gamma_2}\int_0^1\{\mu\left|\frac{\partial u(x,t-\tau\rho)}{\partial t}\right|^2+\xi\left|\frac{\partial v(x,t-\tau\rho)}{\partial t}\right|^2\}d\rho d\Gamma$$

$E_d(t)$ can be rewritten via a change of variable as

$$E_d(t) = \frac{1}{2\tau}\int_t^{t+\tau}\int_{\Gamma_2}\{\mu\left|\frac{\partial u(x,s-\tau)}{\partial t}\right|^2+\xi\left|\frac{\partial v(x,s-\tau)}{\partial t}\right|^2\}d\Gamma ds \tag{29}$$

From (29), we obtain (here and throughout the rest of the paper C is some positive constant different at different occurences)

$$E_d(t) \leq C\int_0^T\int_{\Gamma_2}\{\left|\frac{\partial u(x,s-\tau)}{\partial t}\right|^2+\left|\frac{\partial v(x,s-\tau)}{\partial t}\right|^2\}d\Gamma ds \tag{30}$$

for $0 \leq t+\tau \leq T$ and T large enough.

Step 3.
From Poincaré inequality and Proposition 3.5 of [4], we have for T sufficiently large and for any $\epsilon > 0$

$$\mathcal{E}(0) \leq C \int_0^T \int_{\Gamma_2} \{ \left| \frac{\partial u(x,t)}{\partial \nu} \right|^2 + \left| \frac{\partial u(x,t)}{\partial t} \right|^2 + \left| \frac{\partial v(x,t)}{\partial \nu} \right|^2 + \left| \frac{\partial v(x,t)}{\partial t} \right|^2 \} d\Gamma dt + C\{ \|u\|^2_{L^2(0,T;H^{1/2+\epsilon}(\Omega))} + \|v\|^2_{L^2(0,T;H^{1/2+\epsilon}(\Omega))} \} \tag{31}$$

Inserting the boundary conditions (6) and (7) into (31), we obtain

$$\mathcal{E}(0) \leq C \int_0^T \int_{\Gamma_2} \{ \left| \frac{\partial u(x,t)}{\partial t} \right|^2 + \left| \frac{\partial u(x,t-\tau)}{\partial t} \right|^2 + \left| \frac{\partial v(x,t)}{\partial t} \right|^2 + \left| \frac{\partial v(x,t-\tau)}{\partial t} \right|^2 \} d\Gamma dt + C\{ \|u\|^2_{L^2(0,T;H^{1/2+\epsilon}(\Omega))} + \|v\|^2_{L^2(0,T;H^{1/2+\epsilon}(\Omega))} \} \tag{32}$$

Step 4.
Estimate (30) together with (32) yields

$$E(0) \leq C \int_0^T \int_{\Gamma_2} \{ \left| \frac{\partial u(x,t)}{\partial t} \right|^2 + \left| \frac{\partial u(x,t-\tau)}{\partial t} \right|^2 + \left| \frac{\partial v(x,t)}{\partial t} \right|^2 + \left| \frac{\partial v(x,t-\tau)}{\partial t} \right|^2 \} d\Gamma dt + C\{ \|u\|^2_{L^2(0,T;H^{1/2+\epsilon}(\Omega))} + \|v\|^2_{L^2(0,T;H^{1/2+\epsilon}(\Omega))} \} \tag{33}$$

Step 5.
We drop the lower order terms on the right-hand side of (33) by a compactness-uniqueness argument to obtain

$$E(0) \leq C \int_0^T \int_{\Gamma_2} \{ \left| \frac{\partial u(x,t)}{\partial t} \right|^2 + \left| \frac{\partial u(x,t-\tau)}{\partial t} \right|^2 + \left| \frac{\partial v(x,t)}{\partial t} \right|^2 + \left| \frac{\partial v(x,t-\tau)}{\partial t} \right|^2 \} d\Gamma dt \tag{34}$$

Step 6.
From (28), we have

$$E(T) - E(0) \leq -k \int_0^T \int_{\Gamma_2} \{ \left| \frac{\partial u(x,t)}{\partial t} \right|^2 + \left| \frac{\partial u(x,t-\tau)}{\partial t} \right|^2 + \left| \frac{\partial v(x,t)}{\partial t} \right|^2 + \left| \frac{\partial v(x,t-\tau)}{\partial t} \right|^2 \} d\Gamma dt$$

which together with (34) leads to

$$E(T) \leq \frac{Ck^{-1}}{1 + Ck^{-1}} E(0) \tag{35}$$

The desired conclusion follows now from (35).

References

1. Datko, R.: 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. Komornik, V., Rao, B.: Boundary stabilization of compactly wave equations. Asymptotic Anal. **14**, 339–359 (1997)
4. Lasiecka, I., Triggiani, R.: Carleman estimates and exact boundary controllability for a system of coupled non-conservative second-order hyperbolic equations. In: Lecture Notes in Pure and Applied Mathematics, vol. 188, pp. 215–243. Marcel Dekker, New York (1997)
5. Najafi, M., Sarhangi, G.R., Wang, H.: Stabilizability of coupled wave equations in parallel under various boundary conditions. IEEE Trans. Automat. Control **42**, 1308–1312 (1997)
6. 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)
7. Nicaise, S., Rebiai, S.: Stabilization of the Schrödinger equation with a delay term in boundary feedback or internal feedback. Portugal. Math. **68**, 19–39 (2011)
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)

Model Predictive Control of Temperature and Humidity in Heating, Ventilating and Air Conditioning Systems

Jakob Rehrl[1(✉)], Daniel Schwingshackl[2], and Martin Horn[1]

[1] Institute of Automation and Control,
Graz University of Technology, 8010 Graz, Austria
jakob.rehrl@tugraz.at
[2] Control and Mechatronic Systems, Alpen-Adria-Universität Klagenfurt,
9020 Klagenfurt, Austria

Abstract. The major application of heating, ventilating and air-conditioning (HVAC) systems is the simultaneous control of air temperature and air humidity. Therefore, in a typical industrial HVAC plant the following actuators are available: A cooling coil is used to decrease the air temperature and relative humidity by cooling below the dew point temperature. A steam humidifier is installed to increase the air humidity whereas the air temperature is influenced via a heating coil. Additionally, air temperature and humidity are affected by disturbances acting on the system. These disturbances include outer air temperature and humidity as well as the temperatures of hot water and cool water supply. Consequently, in the setup at hand, a plant with three manipulated inputs, four measurable disturbances and two controlled outputs has to be considered. A predictive control scheme based on a discrete time plant model is presented. The proposed controller computes the manipulated variables by solving an optimization problem at each time step. Simulation and measurement results obtained from an industrial HVAC system are shown.

Keywords: Model predictive control · Heating ventilating and air conditioning systems

1 Introduction

Heating, ventilating and air conditioning (HVAC) systems are used in comfort applications like office space air conditioning. Furthermore, they are also required in industrial applications like inlet air conditioning of engine test benches or for the air conditioning of climate test chambers, e.g. used for automotive tests, see Fig. 1. The two latter fields of application impose stringent specifications on control accuracy.

In order to efficiently operate e.g. a climate chamber, the time required to switch from one temperature/humidity setpoint to another one should be as

The authors would like to thank the company Fischer & Co. Luft- und Klimatechnik in Graz, Austria for their support and for providing the test plant.

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 285–295, 2014.
DOI: 10.1007/978-3-662-45504-3_28

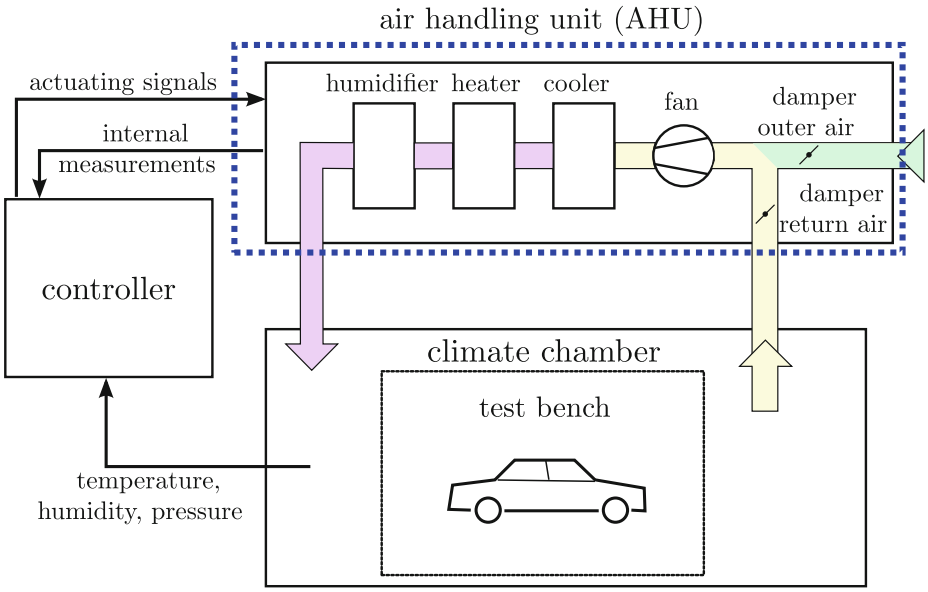


Fig. 1. Sketch of a climate chamber for automotive tests.

short as possible. In Fig. 2, two exemplary setpoint changes are shown: the left one shows poor performance, whereas the right one reveals a much better dynamic behaviour and consequently a shorter time to track new setpoints. In this figure, a given tolerance band for temperature and humidity is indicated. The grey shaded area illustrates the time it takes to meet the tolerance band after a setpoint change.

In the present paper, a concept to control air temperature and air humidity is presented. A systematic approach for controller design is given. The concept is experimentally verified on an industrial test plant and the results are compared to standard control techniques. The paper is structured as follows: Sect. 2 introduces an industrial test plant used to experimentally validate the results. A mathematical model of the plant required for the proposed control concept is given. Section 3 describes the suggested control strategy and its application to the test plant. Section 4 discusses the obtained results and Sect. 5 concludes the paper.

2 Test Plant

In order to verify the proposed concept on an industrial system, the test plant shown in Fig. 3 is available[1]. The plant is capable of increasing and decreasing air temperature as well as air humidity. The core components are heating coils

[1] The test plant was built and is maintained by Company Fischer&Co. Luft- und Klimatechnik in Graz, Austria (http://www.fischer-co.at/).

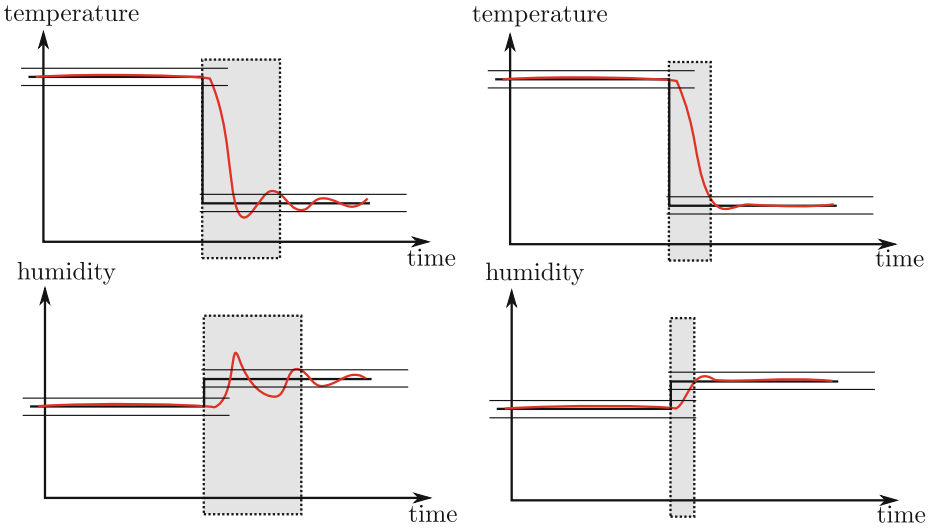


Fig. 2. A setpoint change for temperature and humidity. Poor reference tracking (left hand side) vs. desired reference tracking (right hand side).

(to increase the air temperature), cooling coils (to decrease the air temperature and to decrease the air humidity) and a steam humidifier (to increase the air humidity). Via a fan, the conditioned air can be transported to a neighboring factory building. In the considered plant setup, three actuating signals ($u_1 \ldots$ cooling coil 1, $u_2 \ldots$ heating coil 1, $u_3 \ldots$ steam humidifier) are used. The controlled variables are air temperature y_1 and air humidity y_2 in the supply air duct to the factory building. The air temperature and humidity after the fan, denoted by d_1 and d_2 respectively, are regarded as measurable disturbances. Furthermore, the hot water supply temperature for the heating coil and the cold water supply temperature for the cooling coil are considered as measureable disturbances d_3 and d_4.

2.1 Mathematical Plant Model

Mathematical plant models were derived for the components of the test plant. For the relevant items, the modeling will be described in the following subsections.

Temperature and Humidity Sensor. Temperature and humidity sensors were modeled as first order systems with transfer functions

$$G(s) = \frac{1}{1 + s\,T} \quad , \tag{1}$$

where the individual time constants T were identified from measurements.

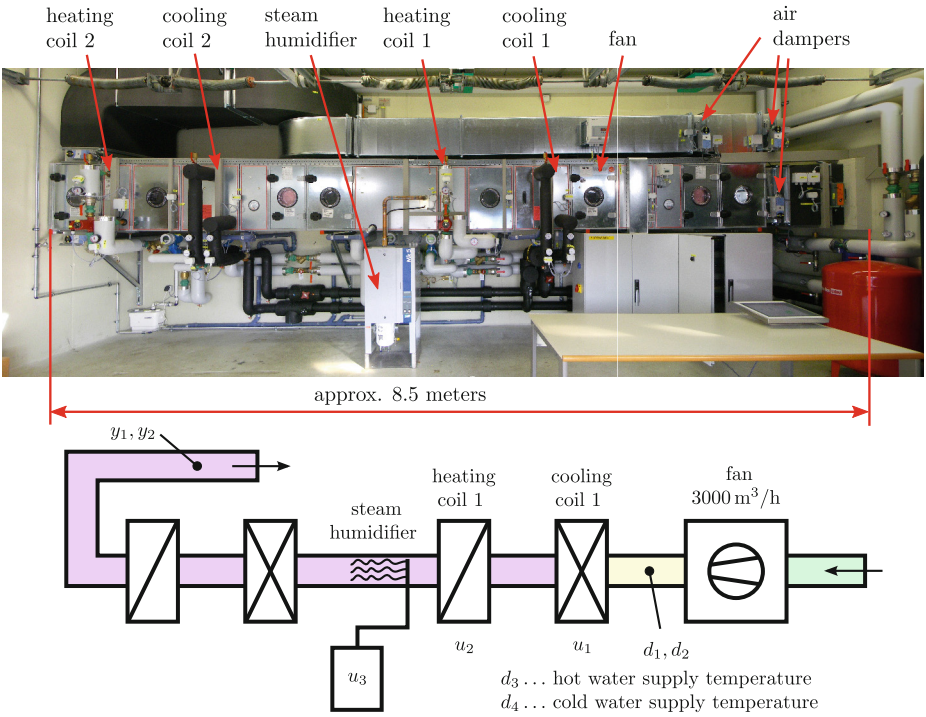


Fig. 3. Picture and schematic representation of the industrial test plant. Outside air enters the plant from the right, is conditioned and then transported to the neighboring factory building.

Heating and Cooling Coil. Since the structure of heating and cooling coil is in principle the same, only one model is required for both. A hot/cold fluid passes through pipes and air circulates around the pipes which leads to - in case of a temperature difference between water and air - a heat transfer. The mathematical model is derived from mass and energy balances. The gained partial differential equations describing temperature of water, pipe and air are converted to ordinary differential equations by segmenting the pipe [1,2]. For one segment, see Fig. 4, the following set of differential equations is obtained[2].

$$\frac{d\vartheta_{\mathrm{p},j}}{dt} = \frac{\alpha_\mathrm{i} A_\mathrm{i}}{m_\mathrm{p} c_\mathrm{p}} \left[\frac{\vartheta^\mathrm{I}_{\mathrm{w},j} + \vartheta^\mathrm{II}_{\mathrm{w},j}}{2} - \vartheta_{\mathrm{p},j} \right] + \frac{\alpha_\mathrm{o} A_\mathrm{o} \Psi_\mathrm{a}}{m_\mathrm{p} c_\mathrm{p} \kappa_\mathrm{a}} (\vartheta^\mathrm{in}_{\mathrm{a},j} - \vartheta_{\mathrm{p},j}) + \frac{\beta A_\mathrm{o} r_\mathrm{v} \Psi_\mathrm{v}}{m_\mathrm{p} c_\mathrm{p} \kappa_\mathrm{v}} (x^\mathrm{in}_{\mathrm{a},j} - x_{\mathrm{p},j}) \tag{2}$$

$$\frac{d\vartheta^\mathrm{I}_{\mathrm{w},j}}{dt} = \frac{2}{T_\mathrm{dw} \Delta\tilde{x}} \left[\vartheta^\mathrm{in}_{\mathrm{w},j} - \frac{\vartheta^\mathrm{I}_{\mathrm{w},j} + \vartheta^\mathrm{II}_{\mathrm{w},j}}{2} \right] + \frac{\alpha_\mathrm{i} A_\mathrm{i}}{m_\mathrm{w} c_\mathrm{w}} (\vartheta_{\mathrm{p},j} - \vartheta^\mathrm{I}_{\mathrm{w},j}) \tag{3}$$

$$\frac{d\vartheta^\mathrm{II}_{\mathrm{w},j}}{dt} = \frac{2}{T_\mathrm{dw} \Delta\tilde{x}} (\vartheta^\mathrm{I}_{\mathrm{w},j} - \vartheta^\mathrm{II}_{\mathrm{w},j}) + \frac{\alpha_\mathrm{i} A_\mathrm{i}}{m_\mathrm{w} c_\mathrm{w}} (\vartheta_{\mathrm{p},j} - \vartheta^\mathrm{II}_{\mathrm{w},j}) \tag{4}$$

[2] A description of the used variables can be found in the nomenclature at the end of the paper.

$$\vartheta_{\mathrm{w},j}^{\mathrm{out}} = 1.5\vartheta_{\mathrm{w},j}^{\mathrm{II}} - 0.5\vartheta_{\mathrm{w},j}^{\mathrm{I}} \tag{5}$$

$$\vartheta_{\mathrm{a},j}^{\mathrm{out}} = e^{-\kappa_{\mathrm{a}}}\vartheta_{\mathrm{a},j}^{\mathrm{in}} + (1 - e^{-\kappa_{\mathrm{a}}})\vartheta_{\mathrm{p},j} \tag{6}$$

$$x_{\mathrm{a},j}^{\mathrm{out}} = e^{-\kappa_{\mathrm{v}}}x_{\mathrm{a},j}^{\mathrm{in}} + (1 - e^{-\kappa_{\mathrm{v}}})x_{\mathrm{p},j} \tag{7}$$

The temperature ϑ_p of the pipe is influenced by the heat transfer from water to pipe, by the air inlet temperature and the inlet humidity, see (2). The water temperatures $\vartheta_{\mathrm{w},j}^{\mathrm{I}}$ and $\vartheta_{\mathrm{w},j}^{\mathrm{II}}$ depend on the inlet water temperature $\vartheta_{\mathrm{w},j}^{\mathrm{in}}$ and the pipe temperature ϑ_p. The coefficients in (2)–(4) are given by heat transfer properties. The outlet air temperature $\vartheta_{\mathrm{a},j}^{\mathrm{out}}$ as well as the outlet air humidity $x_{\mathrm{a},j}^{\mathrm{out}}$ are computed from the respective values at the segments inlet and at the pipe, see (6) and (7). The computation of the weighting factors is given in the nomenclature and can be found in [1].

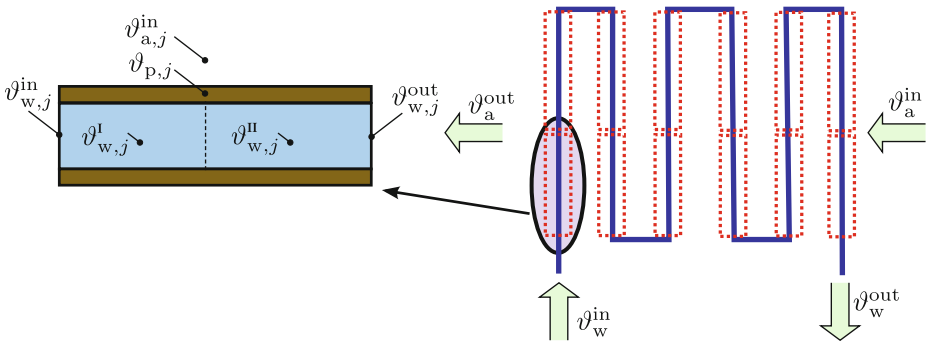


Fig. 4. Structure of a heating/cooling coil and sketch of one pipe segment.

Hydraulics. The structure of the hydraulics for heating coil and cooling coil differ. The heating coil is operated with (almost) constant water mass flow, its heating power is varied by adjusting the mixing ratio of (hot) supply water with (cold) return water. In contrast to this operating mode, the water mass flow of the cooling coil is varied in order to set the cooling power. For the hydraulic system of the heating coil, a static curve which relates a valve position to a mixing ratio is used for modeling. In case of the cooling coil, a curve that relates the valve position to a water mass flow is used. Both relationships were obtained from measurements and valve data. From the valve data and pipe resistance values, the shape of the respective static curve was obtained. Via measurements, the scaling of the static curve was adopted to the plant hydraulics.

3 Control Concept

Model predictive control (MPC) [3,4] has been increasingly applied to HVAC systems in recent years, see e.g. [5–8]. The sampling times of HVAC systems, which are typically in the range of several seconds, make the online solution of

the optimization problem possible, even with limited computing power. Furthermore, multi-input-multi-output systems with constraints on the actuating signals are handled naturally with MPC. The block diagram of the proposed concept is depicted in Fig. 5. The model predictive controller utilizes a linear plant model which is updated each sampling instant. The linear model is obtained from a model constructed via the local linear model tree (LoLiMoT) algorithm [9–11]. This model will be referred to as LoLiMoT-model in the following and its generation will be illustrated in the next section. The computation of the linear state space model is addressed in Sect. 3.2.

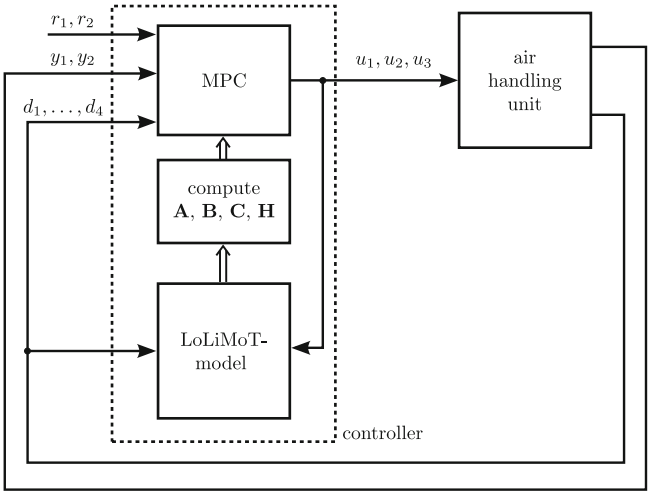


Fig. 5. Block diagram of the proposed control concept.

3.1 LoLiMoT-Model of the Plant

The idea of the LoLiMoT algorithm is to approximate a nonlinear system via several locally affine models. Input and output data of the system to be modeled, in the following referred to as identification data, is required by the algorithm to compute the local models parameters w and the validity range of the local models. The output of one local model is computed from the n previous plant inputs u and outputs y via its individual difference equation. The order of the local models is given by n. The output y_k of the LoLiMoT-model at time instant k is computed via the weighted sum of the locally affine models outputs, i.e.

$$y_k = \sum_{l=1}^{M} \left(w_{l0} + \sum_{i=1}^{n} \left[w_{li}^{y} y_{k-i} + \sum_{j=1}^{m} w_{li}^{u_j} u_{j,k-i} \right] \right) \Phi_l(\mathbf{u^*}_k), \tag{8}$$

$$\text{where} \quad \mathbf{u^*}_k = [u_{1,k-1}\ u_{1,k-2} \ldots u_{1,k-n}\ u_{2,k-1} \ldots u_{2,k-n} \ldots u_{m,k-1} \ldots u_{m,k-n}\ y_{k-1}\ y_{k-2} \ldots y_{k-n}]^T \ . \tag{9}$$

The number of local models is denoted by M. The number of model inputs is m. In the present paper, $M = 20$, $n = 1$ and $m = 7$ holds. The weighting functions $\Phi(\mathbf{u^*}_k)$ corresponding to the local models are normalized Gaussian functions. They depend on the previous inputs and outputs of the LoLiMoT-model which are collected in the vector $\mathbf{u}^*$.

In a first attempt, the identification data is directly derived from measurements at the test plant. This approach shows a severe problem: the disturbances d_1 to d_4 cannot be excited arbitrarily. Consequently, the data available for identification is inappropriate. To tackle the mentioned obstacle, the identification data was generated via the mathematical plant model given in Sect. 2.1. Via this method, sufficiently long and sufficiently excited identification signals can be generated. In Fig. 6, a comparison of test plant measurements versus the mathematical model output is shown. The dynamic behaviour is captured very well. The temperature offset will be compensated by the controller. The LoLiMoT-model output compared to the test plant measurements is depicted in Fig. 6 on the right hand side. A good accordance of the model output with the test plant measurements is given.

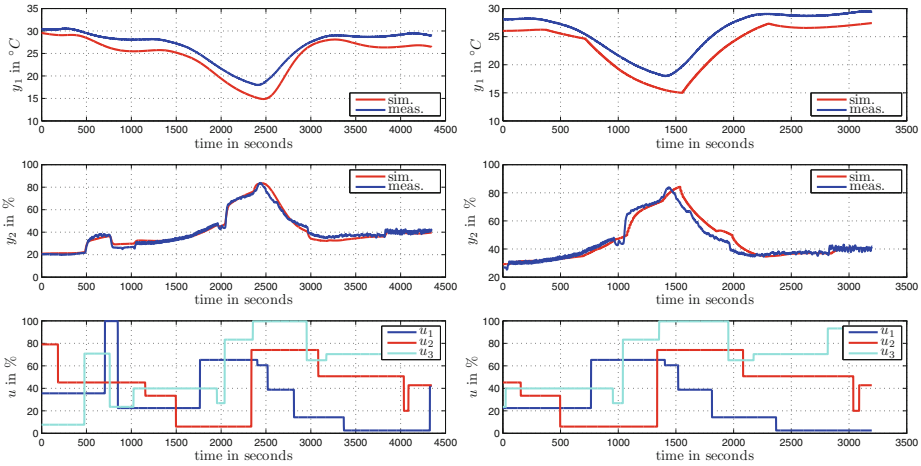


Fig. 6. Comparison: test plant measurements vs. mathematical model output (left) and test plant measurement vs. LoLiMoT-model output (right).

3.2 Computation of the Linear State Space Model

For the proposed control concept, the plant parameters **A**, **B**, **H** and **C** of a linear state space model are required. These parameters are updated at each sampling instant and are obtained from the LoLiMoT-model. The observability canonical form is chosen, the parameters are computed from the coefficients given in (8). A detailed description is omitted due to space limitations, it can be found in [12].

3.3 Model Predictive Controller

The proposed control concept relies on a *linear* model predictive controller of the form[3]

$$\min_{\mathbf{u}_{k+i|k}} \sum_{i=0}^{n_\mathrm{p}-1} (\mathbf{r}_{k+i|k} - \hat{\mathbf{y}}_{k+i|k})^T \mathbf{Q} (\mathbf{r}_{k+i|k} - \hat{\mathbf{y}}_{k+i|k}) + \\ + \Delta \mathbf{u}_{k+i|k}^T \mathbf{R}_1 \Delta \mathbf{u}_{k+i|k} + \mathbf{u}_{k+i|k}^T \mathbf{R}_2 \mathbf{u}_{k+i|k} \tag{10}$$

$$s.t.$$

$$\mathbf{u}_{min} \leq \mathbf{u}_{k+i|k} \leq \mathbf{u}_{max} \tag{11}$$

$$\Delta \mathbf{u}_{min} \leq \Delta \mathbf{u}_{k+i|k} \leq \Delta \mathbf{u}_{max} \tag{12}$$

$$\Delta \mathbf{u}_{k+i|k} = \mathbf{u}_{k+i|k} - \mathbf{u}_{k+i-1|k} \tag{13}$$

$$\Delta \mathbf{u}_{k+i|k} = 0 \quad \forall i > n_\mathrm{c} \tag{14}$$

$$\mathbf{x}_{k+i+1|k} = \mathbf{A}_k \mathbf{x}_{k+i|k} + \mathbf{B}_k \mathbf{u}_{k+i|k} + \mathbf{H}_k \mathbf{d}_{k+i|k} \tag{15}$$

$$\hat{\mathbf{y}}_{k+i|k} = \mathbf{C}_k \mathbf{x}_{k+i|k} \tag{16}$$

Deviations of the predicted plant output $\hat{\mathbf{y}}$ from the reference $\mathbf{r}$ are penalized by the matrix $\mathbf{Q}$ along the prediction horizon n_p. The choice of the prediction horizon was motivated by the system dynamics. From step experiments, the dominant time constant was determined and the horizon was chosen to cover approximately 5 time constants. With this setting, extensive experimental validation was performed on the industrial system with the proposed MPC-LoLiMoT-scheme. These experiments showed that the controlled variables converged to their respective reference values. The predicted output $\hat{\mathbf{y}}_{k+i|k}$ is corrected by the difference between measurement and model output at time k, i.e. $\hat{\mathbf{y}}_k - \mathbf{y}_k$. Constraints on the actuating signal $\mathbf{u}$ are given by (11). The rate of change of the actuating signal $\Delta\mathbf{u}$ is limited by $\Delta\mathbf{u}_{min}$ and $\Delta\mathbf{u}_{max}$, see (12)–(13). Furthermore, the actuating signal is supposed to remain constant for $i > n_c$, where n_c is the control horizon, see (14). Constraints due to the plant model are represented by (15) and (16). The actuating signal as well as the actuating signals rate of change are penalized via $\mathbf{R}_2$ and $\mathbf{R}_1$ respectively.

4 Discussion

In Fig. 7, measurements obtained at the test plant are presented. The diagram on the left hand side outlines the capability of the proposed controller to track reference step signals. The short settling time demonstrates the performance of the MPC/LoLiMoT combination. Actuating signal limits concerning amplitude (limited to the range 0–100 %) and rate (limited to 28.57 % per 10 s) are

[3] The nomenclature $k+i|k$ denotes the prediction of a variable at time instant $k+i$, provided measurement data is available up to time instant k.

accounted for by (11) and (12). The heater is at the lower limit until 1000 s, i.e. constraint (11) is active for the heater during this period. At time equal to 1000 s, the rate limitations (12) are active for the heater and for the humidifier. During the experiment, at least one of the actuators is at its lower limit most of the time. In the diagram on the right hand side, a comparision to a classical PI approach is shown. For this measurement, the steam humidifier was deactivated, and instead of cooling coil 1 and heating coil 1, cooling coil 2 and heating coil 2 were selected as actuators. In the PI-strategy, the cooler was used to control the temperature, the heater was used to control the humidity. Two seperate PI-controllers were tuned by a company specialized to HVAC control. In the comparison diagram, the proposed strategy clearly outperforms the PI strategy. Especially, the temperature can be kept at the setpoint very accurately (notice the small deviation of less than $0.5\,°C$ from the reference) with the proposed strategy, whereas the PI-strategy shows control errors above $1.5\,°C$. The tracking performance regarding the humidity is similar for both approaches with slight advantages for the proposed concept. The humidity remains in a $\pm 5\,\%$ tolerance band from approximately 100 s after the step signal.

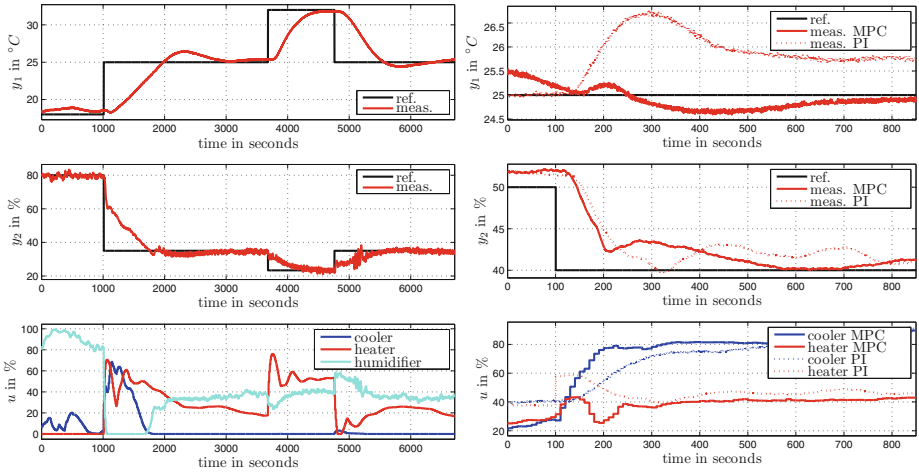


Fig. 7. Measurement results. Left: tracking of several reference steps. Right: Comparison to a conventional PI-controller.

5 Conclusion

In the present paper, a control technique is presented which relies on a *linear* MPC formulation. To deal with nonlinearities of the plant, the parameters of the linear model are updated each sampling instant. A plant model obtained via the LoLiMoT algorithm forms the basis for the creation of the linear state space model. For the presented application consisting of 3 actuators and 2 controlled variables, the presented concept naturally handles the choice of the actuators.

Limitations of the actuators (e.g. limited valve travel in the range of 0 to 100 % and limited slew rate) are handled by the model predictive controller. With conventional schemes, e.g. separate PI-controllers, an additional logic to switch between the destinct controllers has to be implemented. This is not necessary with the proposed concept, it offers a straight forward, systematic approach to design controllers for HVAC systems.

Nomenclature

A_i ... inner pipe surface
A_o ... outer pipe surface
c_a ... specific heat capacity of air
c_p ... specific heat capacity of the pipe
c_w ... specific heat capacity of water
$\dot{m}_\mathrm{a}$ ... air mass flow
m_p ... pipe mass
m_w ... water mass
n_c ... control horizon
n_p ... prediction horizon
r_v ... evaporation heat of water
T_dw ... time it takes the water to pass the coil
T ... time constant
x ... air humidity in kg water per kg air
$\Delta\tilde{x}$ ... normalized length of one pipe segment
α_i ... inner heat transfer coefficient water-pipe
α_o ... outer heat transfer coefficient pipe-air
β ... mass transfer coefficient
ϑ ... temperature, index a: air, index p: pipe, index w: water
κ_a ... $\frac{\alpha_\mathrm{o} A_\mathrm{o}}{\dot{m}_\mathrm{a} c_\mathrm{a}}$
κ_v ... $\frac{\beta A_\mathrm{o}}{\dot{m}_\mathrm{a}}$
Ψ_a ... $1 - e^{\kappa_\mathrm{a}}$
Ψ_v ... $1 - e^{\kappa_\mathrm{v}}$

References

1. Wiening, W.: Zur Modellbildung, Regelung und Steuerung von Wärmeübertragern zum Heizen und Kühlen von Luft. Fortschritt-Berichte VDI Reihe 8 Nr. 128. VDI-Verlag, Düsseldorf (1987)
2. Rehrl, J.: Modeling, Simulation and Control of complex Heating, Ventilating and Air Conditioning (HVAC) Systems. Ph.D. thesis, Alpen-Adria-Universität Klagenfurt (2011)
3. Maciejowski, J.M.: Predictive Control with Constraints. Pearson, London (2002)
4. Camacho, E.F., Bordons, C.: Model Predictive Control, 2nd edn. Springer, New York (2007)
5. Aswani, A., Master, N., Taneja, J., Culler, D., Tomlin, C.: Reducing transient and steady state electricity consumption in HVAC using learning-based model-predictive control. Proc. IEEE **100**(1), 240–253 (2012)
6. Aswani, A., Master, N., Taneja, J., Krioukov, A., Culler, D., Tomlin, C.: Energy-efficient building HVAC control using hybrid system LBMPC. In: 4th IFAC Nonlinear Model Predictive Control Conference, pp. 496–501 (2012)
7. Ma, Y., Kelman, A., Daly, A., Borrelli, F.: Predictive control for energy efficient buildings with thermal storage: modeling, simulation, and experiments. IEEE Control Syst. **32**(1), 44–64 (2012)

8. Oldewurtel, F., Ulbig, A., Parisio, A., Andersson, G., Morari, M.: Reducing peak electricity demand in building climate control using real-time pricing and model predictive control. In: 49th IEEE Conference on Decision and Control (CDC), pp. 1927–1932 (2010)
9. Nelles, O.: Nonlinear System Identification. Springer, New York (2010)
10. Nelles, O.: LOLIMOT - Lokale, lineare Modelle zur Identifikation nichtlinearer, dynamischer Systeme. at - Automatisierungstechnik **45**, 163–174 (1997)
11. Hecker, O., Nelles, O., Moseler, O.: Nonlinear system identification and predictive control of a heat exchanger based on local linear fuzzy models. In: Proceedings of the American Control Conference, vol. 5, pp. 3294–3298, June 1997
12. Schwingshackl, D., Rehrl, J., Horn, M.: Model predictive control of a HVAC system based on the LoLiMoT algorithm. In: European Control Conference (ECC), pp. 4328–4333 (2013)

Regularization of Linear-Quadratic Control Problems with L^1-Control Cost

Christopher Schneider(✉) and Walter Alt

Fakultät Für Mathematik Und Informatik,
Friedrich-Schiller-Universität, 07740 Jena, Germany
{christopher.schneider,walter.alt}@uni-jena.de

Abstract. We analyze L^2-regularization of a class of linear-quadratic optimal control problems with an additional L^1-control cost depending on a parameter β. To deal with this nonsmooth problem we use an augmentation approach known from linear programming in which the number of control variables is doubled. It is shown that if the optimal control for a given $\beta^* \geq 0$ is bang-zero-bang, the solutions are continuous functions of the parameter β and the regularization parameter α. Moreover we derive error estimates for Euler discretization.

Keywords: Optimal control · Bang-bang control · L^1-minimization · Nonsmooth analysis · Regularization · Discretization

1 Introduction

The regularization of optimal control problems by a L^2-term $\frac{\alpha}{2}\|u\|_{L^2}^2$ is often used in order to get a smoother optimal control. In this cases α can be viewed as a regularization parameter and one is interested in the question how the solutions depend on this parameter. For the special case that the control variable appears linearly in the control problem and the optimal control without regularization ($\alpha = 0$) has bang-bang structure this question has been investigated in Deckelnick/Hinze [1] for a class of elliptic control problems and in Alt/Seydenschwanz [2] for a general class of linear-quadratic control problems governed by ordinary differential equations.

Maurer/Vossen [3] investigate first order necessary and second order sufficient optimality conditions for a class of nonlinear control problems involving a L^1-term in the cost functional, where the parameter β is kept fixed. They also propose some numerical algorithms for the solution of such problems. Sakawa [4] also considers a special numerical algorithm for a fixed parameter $\beta > 0$. Stadler [5] and Casas et al. [6,7] investigate classes of elliptic control problems with a L^1-term in the cost functional, which is interpreted as a regularization term. They derive results on the dependence of the solutions on the parameter β and error estimates for discretizations, but an additional L^2-regularization term with fixed parameter α is used in order to get smoother solutions. In Wachsmuth/Wachsmuth [8] the dependence of solutions of a class

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 296–305, 2014.
DOI: 10.1007/978-3-662-45504-3_29

of elliptic control problems on the regularization parameter α is studied while the parameter β is kept fix.

Results for the dependence of the solutions on the parameter β and error estimates for discretizations for a general class of linear-quadratic control problems governed by ordinary differential equations have been recently derived in [9]. In the present paper, we investigate the regularization of such control problems and the dependence of solutions on the parameter β and the regularization parameter α assuming that for a fixed parameter β^* the corresponding optimal control is of bang-zero-bang type.

2 Problem Formulation

With $X = X_1 \times X_2$, $X_1 = W^1_\infty(0, t_f; \mathbb{R}^n)$, $X_2 = L^\infty(0, t_f; \mathbb{R}^m)$, we consider the following family of L^2-regularized linear-quadratic control problems with L^1-control cost depending on the parameters $\alpha \geq 0$ and $\beta \geq 0$:

$$\begin{aligned} \min_{(x,u)\in X} \quad & f_{\alpha,\beta}(x,u) \\ \text{s.t.} \quad & \dot{x}(t) = A(t)x(t) + B(t)u(t) && \text{a.e. on } [0,t_f]\,, \\ & x(0) = a\,, \\ & u(t) \in U && \text{a.e. on } [0,t_f]\,, \end{aligned} \tag{$\mathrm{PQ}_{\alpha,\beta}$}$$

where $f_{\alpha,\beta}$ is a linear-quadratic cost functional with an additional nonsmooth L^1-term defined by

$$\begin{aligned} f_{\alpha,\beta}(x,u) = {} & \frac{1}{2}x(t_f)^\mathsf{T} Q x(t_f) + q^\mathsf{T} x(t_f) \\ & + \int_0^{t_f} \frac{1}{2}x(t)^\mathsf{T} W(t) x(t) + w(t)^\mathsf{T} x(t) + r(t)^\mathsf{T} u(t)\,\mathrm{d}t \\ & + \beta \|u\|_{L^1} + \frac{\alpha}{2}\|u\|^2_{L^2}\,. \end{aligned}$$

Here, $u(t) \in \mathbb{R}^m$ is the control, and $x(t) \in \mathbb{R}^n$ is the state of the system at time t, where $t \in [0, t_f]$. Further $Q \in \mathbb{R}^{n\times n}$ is a symmetric and positive semidefinite matrix, $q \in \mathbb{R}^n$, and the functions $W\colon [0,t_f] \to \mathbb{R}^{n\times n}$, $w\colon [0,t_f] \to \mathbb{R}^n$, $r\colon [0,t_f] \to \mathbb{R}^m$, $A\colon [0,t_f] \to \mathbb{R}^{n\times n}$, and $B\colon [0,t_f] \to \mathbb{R}^{n\times m}$ are Lipschitz continuous. The matrices $W(t)$ are assumed to be symmetric and positive semidefinite, and the set $U \in \mathbb{R}^m$ is defined by lower and upper bounds, i.e.

$$U = \{u \in \mathbb{R}^m \mid b_\ell \leq u \leq b_u\}$$

with $b_\ell, b_u \in \mathbb{R}^m$, $b_\ell < b_u$, where all inequalities are to be understood componentwise.

While the regularization term $\frac{\alpha}{2}\|u\|^2_{L^2}$ leads to a smooth optimal control for $\alpha > 0$ the term $\beta\|u\|_{L^1}$ may be interpreted as both a regularization or some (nonsmooth) L^1-control cost. We are interested in the behavior of a solution $u^{\alpha,\beta}$ of Problem ($\mathrm{PQ}_{\alpha,\beta}$) depending on both parameters α and β.

3 Optimality Conditions

We denote by

$$\mathcal{U} = \{u \in X_2 \mid u(t) \in U \text{ a.e. on } [0, t_f]\}$$

the set of admissible controls, and by

$$\mathcal{F} = \{(x, u) \in X \mid u \in \mathcal{U},\ \dot{x}(t) = A(t)x(t) + B(t)u(t) \text{ a.e. on } [0, t_f],\ x(0) = a\}$$

the feasible set of $(\mathrm{PQ}_{\alpha,\beta})$. Since $\mathcal{U}$ is nonempty, the feasible set $\mathcal{F}$ is nonempty, too. And since $\mathcal{U}$ is bounded, it follows that $\dot{x}$ is bounded for any feasible pair $(x, u) \in \mathcal{F}$, and therefore $\mathcal{F} \subset X$. Moreover, there is some constant c such that $\|x\|_{1,\infty} \le c\,\|u\|_{L^\infty}$ for any solution x of the system equation, which implies that $\mathcal{F}$ is bounded.

A feasible pair $(x^{\alpha,\beta}, u^{\alpha,\beta}) \in \mathcal{F}$ is called a *minimizer* for Problem $(\mathrm{PQ}_{\alpha,\beta})$ if $f_{\alpha,\beta}(x^{\alpha,\beta}, u^{\alpha,\beta}) \le f_{\alpha,\beta}(x, u)$ for all $(x, u) \in \mathcal{F}$. Since the feasible set $\mathcal{F}$ is nonempty, closed, convex and bounded, and the cost functional is convex and continuous, a minimizer $(x^{\alpha,\beta}, u^{\alpha,\beta}) \in W_2^1(0, t_f; \mathbb{R}^n) \times L^2(0, t_f; \mathbb{R}^m)$ of $(\mathrm{PQ}_{\alpha,\beta})$ exists (see [10, Chap. II, Prop. 1.2]), and since $\mathcal{U}$ is bounded we have $(x^{\alpha,\beta}, u^{\alpha,\beta}) \in X = W_\infty^1(0, t_f; \mathbb{R}^n) \times L^\infty(0, t_f; \mathbb{R}^m)$.

Let $(x^{\alpha,\beta}, u^{\alpha,\beta}) \in \mathcal{F}$ be a minimizer of $(\mathrm{PQ}_{\alpha,\beta})$. Then there exist an element $\gamma^{\alpha,\beta} \in \partial\|u^{\alpha,\beta}\|_{L^1}$ of the subdifferential of $\|u^{\alpha,\beta}\|_{L^1}$ and a function $\lambda^{\alpha,\beta} \in W_\infty^1(0, t_f; \mathbb{R}^n)$ such that the adjoint equation

$$\begin{aligned} -\dot{\lambda}^{\alpha,\beta}(t) &= A(t)^{\mathsf{T}}\lambda^{\alpha,\beta}(t) + W(t)x^{\alpha,\beta}(t) + w(t) \quad \text{a.e. on } [0, t_f]\,, \\ \lambda^{\alpha,\beta}(t_f) &= Qx^{\alpha,\beta}(t_f) + q\,, \end{aligned} \tag{1}$$

and the minimum principle

$$\left[B(t)^{\mathsf{T}}\lambda^{\alpha,\beta}(t) + r(t) + \alpha\, u^{\alpha,\beta}(t) + \beta\,\gamma^{\alpha,\beta}(t)\right]^{\mathsf{T}} \left(u - u^{\alpha,\beta}(t)\right) \ge 0 \quad \forall u \in \mathcal{U} \tag{2}$$

hold a.e. on $[0, t_f]$ (compare e.g. [11, Theorem 10.47] or [3, Sect. 2]).

Remark 1. Since $(\mathrm{PQ}_{\alpha,\beta})$ is a convex optimization problem for all $\alpha \ge 0$ and $\beta \ge 0$, a pair $(x^{\alpha,\beta}, u^{\alpha,\beta}) \in \mathcal{F}$ satisfying the minimum principle (2) and solving the adjoint equation (1) with some functions $\gamma^{\alpha,\beta}$ and $\lambda^{\alpha,\beta}$ is a solution of $(\mathrm{PQ}_{\alpha,\beta})$ (compare [11, Propositon 4.12]).

Provided $\alpha = 0$ we are able to evaluate the minimum principle (2) in more detail (compare [3] and [9]) and obtain

$$u_i^{0,\beta}(t) = \begin{cases} b_{u,i}, & \text{if } \xi_i^\beta(t) < -\beta\,, \\ \text{undetermined} \in]\,0, b_{u,i}]\,, & \text{if } \xi_i^\beta(t) = -\beta\,, \\ 0\,, & \text{if } \xi_i^\beta(t) \in]-\beta, \beta[\,, \\ \text{undetermined} \in [\,b_{\ell,i}, 0[\,, & \text{if } \xi_i^\beta(t) = \beta\,, \\ b_{\ell,i}\,, & \text{if } \xi_i^\beta(t) > \beta\,, \end{cases} \tag{3}$$

where $\xi^\beta(t) := B(t)^\mathsf{T}\lambda^{0,\beta}(t) + r(t)$. If we assume that the set of switching times

$$\mathfrak{M}_i^\beta = \left\{ t \in [0,t_f] \middle| \xi_i^\beta(t) = \beta \text{ or } \xi_i^\beta(t) = -\beta \right\} .$$

is finite, then by (3) the i-th component of the optimal control has a bang-zero-bang structure.

4 Problem Transformation

In common with [3] and [9] we formulate a transformed problem ($\mathrm{TQ}_{\alpha,\beta}$) in order to study the dependence of the optimal control on the parameters α and β. This is a well known augmentation approach from linear programming wherewith we obtain a linear-quadratic control problem with smooth cost functional (see e.g. [12]).

Introducing new controls $v \in \tilde{X}_2 := L^\infty(0,t_f;\mathbb{R}^{2m})$ and using the matrix

$$M := \begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{pmatrix} \in \mathbb{R}^{m\times 2m} \tag{4}$$

we have

$$\begin{aligned} \min_{(x,v)\in X_1\times\tilde{X}_2} \quad & \bar{f}_{\alpha,\beta}(x,v) & & \\ \text{s. t.} \quad & \dot{x}(t) = A(t)x(t) + \mathcal{B}(t)v(t) & \text{a.e. on } [0,t_f]\,, & \qquad (\mathrm{TQ}_{\alpha,\beta}) \\ & x(0) = a\,, & & \\ & v(t) \in V & \text{a.e. on } [0,t_f]\,, & \end{aligned}$$

where $\mathcal{B}(t) := B(t)M$. There are new box constraints for the controls,

$$V := \left\{ v \in \mathbb{R}^{2m} \mid v \geq 0\,,\ v_{2i-1} \leq b_{u,i}\,, v_{2i} \leq -b_{\ell,i}\,,\ i = 1,\dots,m \right\},$$

and $f_{\alpha,\beta}$ is a linear-quadratic cost functional:

$$\begin{aligned} \bar{f}_{\alpha,\beta}(x,v) = {} & \frac{1}{2}x(t_f)^\mathsf{T}Qx(t_f) + q^\mathsf{T}x(t_f) \\ & + \int_0^{t_f} \frac{1}{2}x(t)^\mathsf{T}W(t)x(t) + w(t)^\mathsf{T}x(t) + r(t)^\mathsf{T}Mv(t)\,\mathrm{d}t \\ & + \beta\,\|Mv\|_{L^1} + \frac{\alpha}{2}\,\|Mv\|_{L^2}^2\,. \end{aligned}$$

With the same argumentation as above for Problem ($\mathrm{PQ}_{\alpha,\beta}$) we are able to show that a minimizer of Problem ($\mathrm{TQ}_{\alpha,\beta}$) exists. We denote the set of admissible controls by

$$\mathcal{V} = \left\{ v \in \tilde{X}_2 \mid v(t) \in V \text{ a.e. on } [0,t_f] \right\}$$

and the feasible set of Problem $(\mathrm{TQ}_{\alpha,\beta})$ by $\mathcal{T} \subset X_1 \times \tilde{X}_2$, where

$$\mathcal{T} = \{(x, v) \mid v \in \mathcal{V},\ \dot{x}(t) = A(t)x(t) + \mathcal{B}(t)v(t) \text{ a.e. on } [0, t_f],\ x(0) = a\}.$$

Although Problem $(\mathrm{TQ}_{\alpha,\beta})$ admits controls with components v_{2i-1}, v_{2i} being positive simultaneously, such controls cannot be optimal (see [3, Sect. 4], [9, Sect. 3], [12, p.42etseq.]). Therefore, all optimal controls satisfy

$$v_{2i-1}^{\alpha,\beta}(t) = \max\left\{0, u_i^{\alpha,\beta}(t)\right\}, \quad v_{2i}^{\alpha,\beta}(t) = \max\left\{0, -u_i^{\alpha,\beta}(t)\right\}. \tag{5}$$

The optimality conditions also prove this result. By (5) and $v(t) \geq 0$ we now are able to simplify

$$\|Mv\|_{L^1} = \|v\|_{L^1} = \int_0^{t_f} \sum_{i=1}^{2m} v_i(t)\,\mathrm{d}t \quad \text{and} \quad \|Mv\|_{L^2}^2 = \|v\|_{L^2}^2,$$

which nicely shows, that a L^1- or L^2-regularization of the original problem implies the same regularization of the transformed problem. We finally introduce the minimum principle of Problem $(\mathrm{TQ}_{\alpha,\beta})$

$$\left[\sigma^{\alpha,\beta}\right]^{\mathsf{T}} \left(v - v^{\alpha,\beta}(t)\right) \geq 0 \quad \forall v \in \mathcal{V}, \tag{6}$$

where

$$\sigma^{\alpha,\beta} := M^{\mathsf{T}} \left(B(t)^{\mathsf{T}} \lambda^{\alpha,\beta}(t) + r(t)\right) + \alpha\, v^{\alpha,\beta}(t) + \beta\, e, \tag{7}$$

with $e := (1, \ldots, 1)^{\mathsf{T}} \in \mathbb{R}^{2m}$. The adjoint equation (1) as well as the adjoint variables $\lambda^{\alpha,\beta}$ do not change im comparison to Problem $(\mathrm{PQ}_{\alpha,\beta})$. A detailed discussion of the optimality conditions can be found in [3,9].

5 Uniqueness of Solutions

It is well known that the solution of Problem $(\mathrm{TQ}_{\alpha,\beta})$ is uniquely determined for each $\beta \geq 0$, if $\alpha > 0$ (compare e.g. [13, Satz 3.2.5]). This extends with (5) to Problem $(\mathrm{PQ}_{\alpha,\beta})$.

In the case of $\alpha = 0$ we consider a fixed parameter $\beta^* \geq 0$ and assume that the optimal control v^{0,β^*} of Problem $(\mathrm{TQ}_{0,\beta^*})$ is of bang-bang type which implies an optimal control u^{0,β^*} of bang-zero-bang type for Problem $(\mathrm{PQ}_{0,\beta^*})$ by (5). To ensure this we assume that

(B1) There exists a solution $(x^{0,\beta^*}, v^{0,\beta^*}) \in \mathcal{T}$ of $(\mathrm{TQ}_{0,\beta^*})$ such that the set Σ of zeros of the components of the switching function σ^{0,β^*} defined by (7) is finite and $0, t_f \notin \Sigma$, i.e. $\Sigma = \{s_1, \ldots, s_l\}$ with $0 < s_1 < \ldots < s_l < t_f$.

Let $\mathcal{I}(s_j) := \{1 \leq i \leq 2m \mid \sigma_i^{0,\beta^*}(s_j) = 0\}$ be the set of active indices for the components of the switching function. In order to get stability of the bang-bang structure under perturbations we need an additional assumption (compare [14]):

(B2) The functions B and r are differentiable, $\dot{B}$ and $\dot{r}$ are Lipschitz continuous, and there exists $\bar{\sigma} > 0$ such that

$$\min_{1\le j\le l}\ \min_{i\in\mathcal{I}(s_j)} \left\{ \left|\dot{\sigma}_i^{0,\beta^*}(s_j)\right| \right\} \ge 2\bar{\sigma}\,.$$

Remark 2. Assumption (B2) can be slightly relaxed (see e.g. [9,15]).

The following result is extracted from [14, Proof of Lemma 3.3]. Proofs can also be found in [2,9,15].

Lemma 1. *Let $(x^{0,\beta^*}, v^{0,\beta^*})$ be a minimizer for Problem $(\mathrm{TQ}_{0,\beta^*})$ and let the switching function $\sigma^{0,\beta^*}(t)$ be defined by (7). If Assumptions (B1) and (B2) are satisfied, then there are constants ω, γ, $\bar{\delta} > 0$ independent of β such that for any feasible pair (x, v)*

$$\int_0^{t_f} \sigma^{0,\beta^*}(t)^\mathsf{T} \left(v(t) - v^{0,\beta^*}(t) \right) \mathrm{d}t \ge \omega \left\| v - v^{0,\beta^*} \right\|_{L^1}^2 \tag{8}$$

if $\|v - v^{0,\beta^*}\|_{L^1} \le 2\gamma\bar{\delta}$, and

$$\int_0^{t_f} \sigma^{0,\beta^*}(t)^\mathsf{T} \left(v(t) - v^{0,\beta^*}(t) \right) \mathrm{d}t \ge \omega \left\| v - v^{0,\beta^*} \right\|_{L^1} \tag{9}$$

if $\|v - v^{0,\beta^*}\|_{L^1} \ge 2\gamma\bar{\delta}$.

By the help of standard arguments this result implies uniqueness of the solution of $(\mathrm{TQ}_{0,\beta^*})$ (compare [14, Theorem 2.2]). It follows with (5) that Problem $(\mathrm{PQ}_{0,\beta^*})$ has a unique solution, too.

6 Calmness of Solutions

In this section for $\alpha \ge 0$ and $\beta \ge 0$ we denote by $(x^{\alpha,\beta}, u^{\alpha,\beta})$ and $(x^{\alpha,\beta}, v^{\alpha,\beta})$ the solutions of $(\mathrm{PQ}_{\alpha,\beta})$ and $(\mathrm{TQ}_{\alpha,\beta})$, respectively. We want to study the dependence of solutions on α and β. We derive estimates which show that the solutions as functions of the regularization parameters α and β are *calm at $\alpha = 0$ and $\beta = \beta^*$* (compare Dontchev/Rockafellar [16, Sect. 1C]). For this purpose we combine the results achieved in [2,9].

Theorem 1. *Let* (B1) *and* (B2) *be satisfied for some $\beta^* \ge 0$. Then for any $\alpha \ge 0$ and $\beta \ge 0$ the estimate*

$$\left\| v^{\alpha,\beta} - v^{0,\beta^*} \right\|_{L^1} \le c_1 \left(\alpha + |\beta - \beta^*| \right) \tag{10}$$

holds, where the constant c_1 is independent of α and β.

Proof. We only consider the case $\|v^{\alpha,\beta} - v^{0,\beta^*}\|_{L^1} \leq 2\gamma\bar{\delta}$ and refer to [9] and [2] for the case $\|v^{\alpha,\beta} - v^{0,\beta^*}\|_{L^1} \geq 2\gamma\bar{\delta}$ which can be handled analogously. Since Assumptions (B1) and (B2) are satisfied, for $\alpha, \beta \geq 0$ by (8) we have

$$\int_0^{t_f} \sigma^{0,\beta^*}(t)^\mathsf{T} \left(v^{\alpha,\beta}(t) - v^{0,\beta^*}(t)\right) \mathrm{d}t \geq \omega \left\|v^{\alpha,\beta} - v^{0,\beta^*}\right\|_{L^1}^2 \tag{11}$$

with $\omega > 0$. By the minimum principle (6) we obtain

$$\int_0^{t_f} \sigma^{\alpha,\beta}(t)^\mathsf{T} \left(v^{0,\beta^*}(t) - v^{\alpha,\beta}(t)\right) \mathrm{d}t \geq 0\,. \tag{12}$$

Adding (12) and (11) it follows that

$$\int_0^{t_f} \left(\sigma^{0,\beta^*}(t) - \sigma^{\alpha,\beta}(t)\right)^\mathsf{T} \left(v^{\alpha,\beta}(t) - v^{0,\beta^*}(t)\right) \mathrm{d}t \geq \omega \left\|v^{\alpha,\beta} - v^{0,\beta^*}\right\|_{L^1}^2. \tag{13}$$

Since

$$\sigma^{0,\beta^*}(t) - \sigma^{\alpha,\beta}(t) = \mathcal{B}(t)^\mathsf{T} \left(\lambda^{0,\beta^*}(t) - \lambda^{\alpha,\beta}(t)\right) + (\beta^* - \beta)\, e - \alpha\, v^{\alpha,\beta}(t)\,,$$

and due to the fact that $x^{\alpha,\beta}, x^{0,\beta^*}$ satisfy the system equation, and $\lambda^{\alpha,\beta}, \lambda^{0,\beta^*}$ satisfy the adjoint equation we obtain

$$\begin{aligned}
&\int_0^{t_f} \left[\mathcal{B}(t)^\mathsf{T} \left(\lambda^{0,\beta^*}(t) - \lambda^{\alpha,\beta}(t)\right)\right]^\mathsf{T} \left(v^{\alpha,\beta}(t) - v^{0,\beta^*}(t)\right) \mathrm{d}t \\
&= \left(x^{0,\beta^*}(t_f) - x^{\alpha,\beta}(t_f)\right)^\mathsf{T} Q \left(x^{\alpha,\beta}(t_f) - x^{0,\beta^*}(t_f)\right) \\
&\quad + \int_0^{t_f} \left(x^{0,\beta^*}(t) - x^{\alpha,\beta}(t)\right)^\mathsf{T} W(t) \left(x^{\alpha,\beta}(t) - x^{0,\beta^*}(t)\right) \mathrm{d}t\,.
\end{aligned}$$

Together with (13) this implies

$$\begin{aligned}
&\omega \left\|v^{\alpha,\beta} - v^{0,\beta^*}\right\|_{L^1}^2 + \left(x^{\alpha,\beta}(t_f) - x^{0,\beta^*}(t_f)\right)^\mathsf{T} Q \left(x^{\alpha,\beta}(t_f) - x^{0,\beta^*}(t_f)\right) \\
&\quad + \int_0^{t_f} \left(x^{\alpha,\beta}(t) - x^{0,\beta^*}(t)\right)^\mathsf{T} W(t) \left(x^{\alpha,\beta}(t) - x^{0,\beta^*}(t)\right) \mathrm{d}t \\
&\leq \int_0^{t_f} \left[(\beta^* - \beta)\, e - \alpha\, v^{\alpha,\beta}(t)\right]^\mathsf{T} \left(v^{\alpha,\beta}(t) - v^{0,\beta^*}(t)\right) \mathrm{d}t \\
&\leq |\beta - \beta^*| \left\|v^{\alpha,\beta} - v^{0,\beta^*}\right\|_{L^1} + \alpha \left\|v^{\alpha,\beta}\right\|_{L^\infty} \left\|v^{\alpha,\beta} - v^{0,\beta^*}\right\|_{L^1}.
\end{aligned}$$

Since the matrices Q and $W(t)$, $t \in [0, t_f]$, are assumed to be positive semidefinite and $\omega > 0$, we obtain

$$\omega \left\|v^{\alpha,\beta} - v^{0,\beta^*}\right\|_{L^1}^2 \leq \left(|\beta - \beta^*| + \alpha \left\|v^{\alpha,\beta}\right\|_{L^\infty}\right) \left\|v^{\alpha,\beta} - v^{0,\beta^*}\right\|_{L^1}.$$

We now get (10) with some constant c_1 independent of α and β. □

Remark 3. By Theorem 1 we also obtain estimates for the optimal states

$$\left\|x^{\alpha,\beta} - x^{0,\beta^*}\right\|_{1,1} \leq \bar{c}_1 \left(\alpha + |\beta - \beta^*|\right)$$

and for the optimal controls $u^{\alpha,\beta}$ of the original problem $(\mathrm{PQ}_{\alpha,\beta})$, by using the matrix (4) and the relation (5) between $u^{\alpha,\beta}$ and $v^{\alpha,\beta}$

$$\begin{aligned}\left\|u^{\alpha,\beta} - u^{0,\beta^*}\right\|_{L^1} &= \left\|Mv^{\alpha,\beta} - Mv^{0,\beta^*}\right\|_{L^1} \leq \|M\|_1 \left\|v^{\alpha,\beta} - v^{0,\beta^*}\right\|_{L^1} \\ &\leq c_1 \left(\alpha + |\beta - \beta^*|\right).\end{aligned}$$

If we choose some β in a sufficiently small neighborhood of β^* this result can even be improved.

Theorem 2. *Let* (B1) *and* (B2) *be satisfied for some* $\beta^* \geq 0$. *Then there exist* $\rho > 0$ *and a constant* c_2 *independent of* $\alpha \geq 0$ *and* ρ, *such that for any* $\beta_i \in \mathbb{R}$, $i = 1, 2$, *with* $\beta_i \geq 0$ *and* $|\beta_i - \beta^*| < \rho$ *the estimate*

$$\left\|v^{\alpha,\beta_1} - v^{0,\beta_2}\right\|_{L^1} \leq c_2 \left(\alpha + |\beta_1 - \beta_2|\right) \tag{14}$$

holds.

Proof. We use [9, Theorem 6.3, Remark 10], which proved the local Lipschitz-continuity of the optimal control depending on β, where the constant $\tilde{c}$ is independent of β:

$$\left\|u^{0,\beta_1} - u^{0,\beta_2}\right\|_{L^1} \leq \tilde{c}\,|\beta_1 - \beta_2|\,. \tag{15}$$

In addition to this we are able to extend the result of [2, Theorem 4.1] using the problem transformation introduced in Sect. 4 and obtain

$$\left\|u^{\alpha,\beta_1} - u^{0,\beta_1}\right\|_{L^1} \leq \bar{c}\,\alpha \tag{16}$$

with some constant $\bar{c}$ independent of α. Together (15) and (16) lead to

$$\begin{aligned}\left\|u^{\alpha,\beta_1} - u^{0,\beta_2}\right\|_{L^1} &\leq \left\|u^{\alpha,\beta_1} - u^{0,\beta_1}\right\|_{L^1} + \left\|u^{0,\beta_1} - u^{0,\beta_2}\right\|_{L^1} \\ &\leq \tilde{c}\,\alpha + \bar{c}\,|\beta_1 - \beta_2|\,,\end{aligned}$$

which implies (14). □

7 Discretization

For the numerical solution of Problem $(\mathrm{PQ}_{\alpha,\beta})$ we use the Euler discretization scheme described in [9,15]. Given a natural number N and let $h_N = t_f/N$ be the meshsize, we approximate the cost functional $f_{\alpha,\beta}$ by

$$\begin{aligned}f_{\alpha,\beta,N}(x,u) = &\frac{1}{2}x_N^{\mathsf{T}} Q x_N + q^{\mathsf{T}} x_N + h_N \sum_{i=0}^{N-1} \frac{1}{2} x_i^{\mathsf{T}} W(t_i) x_i + w(t_i)^{\mathsf{T}} x_i + r(t_i)^{\mathsf{T}} u_i \\ &+ h_N \left(\beta \sum_{i=0}^{N-1} \sum_{j=1}^{m} |u_{j,i}| + \frac{\alpha}{2} \sum_{i=0}^{N-1} \sum_{j=1}^{m} u_{j,i}^2 \right),\end{aligned}$$

and Problem $(\mathrm{PQ}_{\alpha,\beta})$ by

$$\begin{aligned} &\min f_{\alpha,\beta,N}(x,u) \\ &\text{s.t. } x_{i+1} = x_i + h_N\left(A(t_i)x_i + B(t_i)u_i\right)\,,\ i = 0,\dots,N-1\,, \qquad (\mathrm{PQ}^{\mathrm{N}}_{\alpha,\beta}) \\ &\qquad x_0 = a\,, \\ &\qquad u_i \in U\,, \qquad i = 0,\dots,N-1\,. \end{aligned}$$

Remark 4. Note that analogously to [9] we solve a transformed discretized problem (compare also Sect. 4) to compute the solution of Problem $(\mathrm{PQ}_{\alpha,\beta})$ numerically.

Theorem 3. *Let $(x^{0,\beta^*}, u^{0,\beta^*})$ be the solution of Problem $(\mathrm{PQ}_{0,\beta^*})$ for which Assumptions* (B1) *and* (B2) *are satisfied. Then, for sufficiently large N, choosing $\alpha = c_\alpha h_N$ and $\beta = \beta^* + c_\beta h_N$ with constants c_α and c_β, any optimal control $u_h^{\alpha,\beta}$ of Problem $(\mathrm{PQ}^{\mathrm{N}}_{\alpha,\beta})$ can be estimated by*

$$\left\|u_h^{\alpha,\beta} - u^{0,\beta^*}\right\|_{L^1} \le c_u h_N\,,$$

where the constant c_u is independent of N.

Proof. Using [17, Theorem 5.2] and [9, Theorem 5.1,Remark 8] we have

$$\begin{aligned} \left\|u_h^{\alpha,\beta} - u^{0,\beta^*}\right\|_{L^1} &\le \left\|u_h^{\alpha,\beta} - u^{0,\beta}\right\|_{L^1} + \left\|u^{0,\beta} - u^{0,\beta^*}\right\|_{L^1} \\ &\le c_\alpha h + \tilde{c}_\beta\,|\beta - \beta^*| \end{aligned}$$

with some constant $\tilde{c}_\beta$ independent of β, which implies the assertion. □

Example 1. (The Rocket Car) We consider the popular example of the rocket car, driving from some starting point to it's destination $(0,0)$.

$$\begin{aligned} &\min \frac{1}{2}\left(x_1(5)^2 + x_2(5)^2\right) + \beta\,\|u\|_{L^1} + \frac{\alpha}{2}\,\|u\|^2_{L^2} \\ &\text{s.t. } \dot{x}_1(t) = x_2(t)\,,\ \dot{x}_2(t) = u(t) \text{ a.e. on } [0,5]\,, \\ &\qquad x_1(0) = 6\,, \quad x_2(0) = 1\,, \\ &\qquad u(t) \in [-1,1] \qquad \text{a.e. on } [0,5]\,. \end{aligned}$$

Table 1 shows numerical results for different meshsizes which confirm the theoretical findings of Theorem 3. To solve the discretized problems we used Ipopt [18].

Table 1. Discretization for different N, $\beta^* = 1$, $\beta = \beta^* + h_N$ and $\alpha = 10\,h_N$.

N	125	250	500	1000	2000	4000
$\|u_h^{\alpha,\beta} - u^{0,\beta^*}\|_{L^1}$	0.2644	0.1344	0.0644	0.0331	0.0177	0.0083
$\frac{\|u_h^{\alpha,\beta} - u^{0,\beta^*}\|_{L^1}}{h_N}$	6.6098	6.7177	6.4409	6.6123	7.0826	6.6752

References

1. 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)
2. Alt, W., Seydenschwanz, M.: Regularization and discretization of linear-quadratic control problems. Control Cybern. **40**(4), 903–920 (2011)
3. Vossen, G., Maurer, H.: On L^1-minimization in optimal control and applications to robotics. Optimal Control Appl. Methods **27**(6), 301–321 (2006)
4. Sakawa, Y.: Trajectory planning of a free-flying robot using the optimal control. Optimal Control Appl. Methods **20**(5), 235–248 (1999)
5. Stadler, G.: Elliptic optimal control problems with L^1-control cost and applications for the placement of control devices. Comput. Optim. Appl. **44**(2), 159–181 (2009)
6. Casas, E., Herzog, R., Wachsmuth, G.: Approximation of sparse controls in semilinear equations by piecewise linear functions. Numer. Math. **122**(4), 645–669 (2012)
7. Casas, E., Herzog, R., Wachsmuth, G.: Optimality conditions and error analysis of semilinear elliptic control problems with L^1-cost functional. SIAM J. Optim. **22**(3), 795–820 (2012)
8. Wachsmuth, G., Wachsmuth, D.: Convergence and regularization results for optimal control problems with sparsity functional. ESAIM Control Optim. Calc. Var. **17**(3), 858–886 (2011)
9. Alt, W., Schneider, C.: Linear-quadratic control problems with L^1-control cost. Optimal Control Appl. Methods (2014). doi:10.1002/oca.2126
10. Ekeland, I., Temam, R.: Convex Analysis and Variational Problems. Studies in Mathematics and its Applications, vol. 1. North-Holland Publishing Company, Amsterdam (1976)
11. Clarke, F.H.: Functional Analysis, Calculus of Variations and Optimal Control. Springer, London (2013)
12. Murty, K.G.: Operations Research: Deterministic Optimization Models. Prentice-Hall, Englewood Cliffs (1995)
13. Alt, W., Schneider, C., Seydenschwanz, M.: EAGLE-STARTHILFE Optimale Steuerung. Theorie und Verfahren, Edition am Gutenbergplatz Leipzig (2013)
14. Felgenhauer, U.: On stability of bang-bang type controls. SIAM J. Control Optim. **41**(6), 1843–1867 (2003)
15. Alt, W., Baier, R., Gerdts, M., Lempio, F.: Error bounds for Euler approximation of linear-quadratic control problems with bang-bang solutions. Numer. Algebra, Control Optim. **2**(3), 547–570 (2012)
16. Dontchev, A.L., Rockafellar, R.T.: Implicit Functions and Solution Mappings. Springer, New York (2009)
17. Seydenschwanz, M.: Improved error estimates for discrete regularization of linear-quadratic control problems with bang-bang solutions (2013) (Submitted)
18. Wächter, A., Biegler, L.T.: On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Math. Program. **106**(1), 25–57 (2006)

Deployment of Sensors According to Quasi-Random and Well Distributed Sequences for Nonparametric Estimation of Spatial Means of Random Fields

Ewa Skubalska-Rafajłowicz and Ewaryst Rafajłowicz(✉)

Institute of Computer Engineering Control and Robotics,
Wroclaw University of Technology, Wroclaw, Poland
{ewa.rafajlowicz,ewaryst.rafajlowicz}@pwr.wroc.pl

Abstract. Our aim is to discuss advantages of quasi-random points (also known as uniformly distributed (UD) points [8]) and their sub-class recently proposed by the authors [17] that are well-distributed (WD) as sensors' positions in estimating the spatial mean. UD and WD s sequences have many interesting properties that are useful both for wireless sensors networks (coverage an and connectivity) and for large area networks such as radiological or environment pollution monitoring stations.

In opposite to most popular parameter estimation approaches, we consider a nonparametric estimator of the spatial mean. We shall prove the estimator convergence in the integrated mean square-error sense.

Keywords: Sensor networks · Quasi-random sequences · Space-filling curves · Nonparametric estimation

1 Introduction

Spatial sampling is a crucial issue for proper estimation of parameters in spatio-temporal dynamical models [9,18] and for estimation of spatial fields (see [14]). For wireless sensor network (WSN) at least three requirements are crucial, namely, efficient energy usage, coverage and connectivity. Here, we concentrate on the last two of them from the view-point of sensors' deployment. Coverage (or information coverage) is the ability of WSN to cover the whole area, assuming that a single sensor has the ability of collecting information from its (usually circular) neighborhood. The connectivity requires that wireless senors can transmit information from one to another and finally to a sink. Both connectivity and coverage require sensors to be evenly placed in the area.

Our aim is to discuss advantages of equidistributed (EQD) (also known as uniformly distributed (UD) or quasi-random points [8]) and their sub-class recently proposed by the authors [17] that are well-distributed (WD). Furthermore, we shall prove that it is possible to construct a nonparametric estimator

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 306–316, 2014.
DOI: 10.1007/978-3-662-45504-3_30

of the mean of (possibly correlated) random field that is based on observations from such points.

We refer the reader to [1] and [3] for surveys on WSN and to [5,10,11,14,19] for an excerpt of approaches recently proposed for sensors placement.

2 Equidistributed Sequences – Good Candidates for Sensors' Sites

Equidistributed sequences, also called uniformly distributed, quasi-random or quasi Monte-Carlo sequences, are well known in the theory of numerical integration (see, e.g., [7,8]). Here, we summarize some of their basic properties, putting emphasis on those, which indicate that they are good candidates for sensors' positions in WSN.

2.1 Definition

Define $I_d = [0, 1]^d$ as d-dimensional unit cube, which is our space for sensors' deployment. Clearly, most WSN are considered for $d = 2$, but it can also be of interest in some applications to consider WSN in 3D space, e.g., for air pollution.

A deterministic sequence $(x_i)_{i=1}^n$ is called EQD sequence in I_d iff

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} g(x_i) = \int_{I_d} g(x)\, dx \tag{1}$$

holds for every g continuous on I_d.

Thus, formally, EQD sequence behave as uniformly distributed random sequences, since (1) mimics a law of large numbers. As we shall see later, EQD sequences are in some sense "more uniform" than uniformly distributed random sequences.

The well known EQD sequences include Corput, Halton, Hammersley, Korobov, Zaremba and Sobol (see, e.g., [6–8]). For our purposes, the following example is important.

Important example – the Weyl sequence is defined as follows:

$$t_i = \text{fractional part}(i\,\theta), \quad i = 1, 2, \ldots,$$

where θ is selected irrational number. E.g., quadratic irrational – $\theta = (\sqrt{5}-1)/2$ behaves quite well in practice. In general, as it is more difficult to approximate θ by rational numbers, then it is better candidate for generating the Weyl sequence.

2.2 Discrepancy as Indicator of WSN Coverage and Connectivity

A discrepancy D_n of EQD sequence is well known measure of its uniformity (see, e.g., [6,7]). D_n discrepancy of $(x_i)_{i=1}^n$ is defined as follows:

$$D_n = \sup_{A \subset I_d} \left| \mu_d(A) - \frac{N_n(A)}{n} \right|, \tag{2}$$

where A is any parallelepiped and supremum is taken with respect to all such $A \subset I_d$, $\mu_d(A)$ is d-dimensional Lebesgue measure (the area or volume) of A, while $N_n(A)$ is number of x_i's in A.

As in the classical Monte-Carlo method, one can expect that for fairly spaced sensors $\mu_d(A)$ is close to the fraction of sensors in A. For interpreting (8) assume that the supremum is attained for a certain set $\hat{A}_n \subset I_d$. Then, $\hat{A}_n$ is the set that is most unevenly covered by sensors, i.e., it contains too small or too large number of sensors in comparison to other areas.

For this reason, in our opinion, D_n is a a good measure for connectivity and coverage of a sensors' net. It is, however, not easy to find $\hat{A}_n$ numerically and for this reason in the theory of numerical integration a simplified, but still useful, version of discrepancy, called D_n^* discrepancy, is more frequently used.

Let $\Pi(x)$ be the parallelepiped in I_d with vertices in $(0, \ldots 0)$ and x. D_n^* discrepancy of $(x_i)_{i=1}^n$ is defined as follows:

$$D_n^* = \sup_{x \in I_d} \left| \mu_d(\Pi(x)) - \frac{N_n(x)}{n} \right|, \tag{3}$$

where $N_n(x)$ is the number of x_i's in $\Pi(x)$.

Notice that D_n^* is an analog to Kolmogorov-Smirnoff statistics for testing the uniformity of a distribution. D_n^* can be efficiently calculated for a given set of points in $1D$, $2D$. Furthermore, it can be proved (see [7]) that

$$D_n^* \leq D_n \leq 2^d D_n^*. \tag{4}$$

For "good" known EQD sequences:

$$D_n^* = \frac{\log^d(n)}{n}.$$

This is much better than for "usual" uniformly distributed random variables, for which

$$D_n^* \sim \frac{1}{\sqrt{n}}.$$

The same order $1/\sqrt{n}$ is obtained when equidistant grid is considered in R^2.

3 Basic Properties of Space-Filling Curves

Our idea is to transform a sequence of one dimensional EQD sequence by a space-filling curve (SFC) in order to obtain multidimensional EQD sequence with good properties. Note that a similar construction has been already proposed and used in the theory of numerical integration, but the transformed sequence was equidistant, i.e., i/n, $i = 1, 2, \ldots, n$.

Below, we summarize known properties of space filling curves that are either directly used in the rest of the paper or indicate why we can expect that our construction leads to sensors' site with good coverage and connectivity.

Space-filling curve is a mapping $\Phi : [0, 1] \overset{onto}{\rightarrow} I_d$ such that

- $\Phi(t)$ a continuous function in $I_1 = [0, 1]$
- maps $I_1 = [0, 1]$ onto I_d – d-dim. cube.

The Hilbert, Peano and Sierpinski are well known examples of SFCs [15]. All these curves:

– preserve areas and neighbors (see SFC (2), SFC (3) below),
– fill the space uniformly and as densely as desired.
– can be generated recursively and efficiently.

For the Peano, Hilbert and Sierpiński curves the following properties hold:

SFC (1) $\forall g : I_d \rightarrow R$, g - continuous

$$\int_{I_d} g(x)\, dx = \int_0^1 g(\Phi(t))\, dt, \tag{5}$$

where $x = [x^{(1)}, x^{(2)}, \ldots, x^{(d)}]$

SFC (2) Hölder continuity:

$$\exists_{C_\Phi > 0}\ ||\Phi(t) - \Phi(t')|| \leq C_\Phi |t - t'|^{1/d}, \tag{6}$$

where $||.||$ is the Euclidean norm in R^d.

It is well known that Φ does not have the inverse. However, for any Borel set $A \subset I_d$ we can define $\Phi^{-1}(A)$ as preimage of A.

SFC (3) Φ preserves the Lebesgue measure in the sense that

$$\forall_{A \subset I_d},\ \ \mu_d(A) = \mu_1(\Phi^{-1}(A)), \tag{7}$$

where μ_1, μ_d are the Lebesgue measures in R_1, R_d, respectively. In other words, volume of Borel set A in I^d is equal to the length of $\Phi^{-1}(A)$ in I_1.

We stress that the Peano, Hilbert and Sierpiński curves can be approximated efficiently, using $O\left(\lceil \frac{d}{\varepsilon} \rceil\right)$ arithmetic operations, where $\varepsilon > 0$ is the approximation accuracy (see [2], [16]). Notice that for our purpose we shall calculate $\Phi(t)$ once for $t = t_1, t_2, \ldots, t_n$. Approximation of the Hilbert SFC is shown in Fig. 1.

4 Proposed EQD Sequences

The following algorithm generates EQD sequences that can be used as sensors' positions.

Algorithm 1

Step (1) Generate EQD sequence in [0, 1] by the Weyl method: $t_i = \text{frac}(i\,\theta),\ i = 1, 2, \ldots, n$, where θ is irrational number, e.g., $\theta = (\sqrt{5} - 1)/2$.

Step (2*) Sort t_i's and get $t_{(1)} < t_{(2)} < \ldots < t_{(n)}$. This step is optional, it serves mainly for theoretical purposes, but it can also be used to determine ordering of senors along SFC, which – in turn – can be used to indicate the ordering of information transmission between sensors.

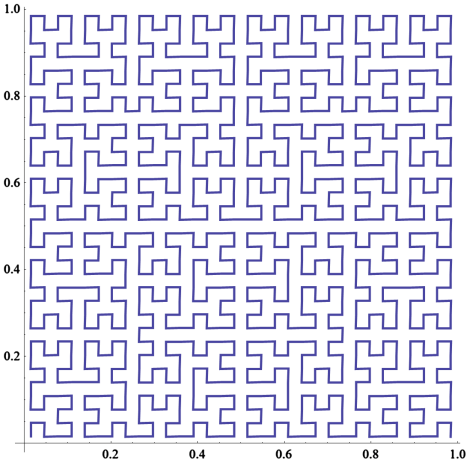


Fig. 1. Approximation of the Hilbert SFC using 1024 points

Step (3) Select SFC and generate x_i's as follows: $x_i = \Phi(t_{(i)})$, $i = 1, 2, \ldots, n$.

Below we state properties of the above generated sequences that justify their usefulness as sensors' positions.

Algorithm 1 generates sequences that are extendable, i.e., one can add points without recalculating positions of earlier sites. This property is not shared by many other methods of generating multidimensional EQD sequences.

A deterministic sequence $(x_i)_{i=1}^n$ is called well distributed (WD) in I_d iff

$$\lim_{n\to\infty} n^{-1} \sum_{i=p}^{p+n} g(x_i) = \int_{I_d} g(x)\, dx$$

holds uniformly in p, for every $g \in C(I_d)$, i.e., continuous on I_d.

It is known that Weyl seq. $t_i = frac(i\,\theta)$ is WD. As far as we know, multidimensional, extendable WD sequences are not known. The only exception is the above proposed sequence.

Theorem: If θ irrational and SFC has the property SFC (1), then x_i's generated by Algorithm 1 are not only EQD but also well distributed.

Proof – EQD property. $\forall g \in C(I_d)$ we have:

$$n^{-1} \sum_{i=1}^{n} g(x_i) = n^{-1} \sum_{i=1}^{n} g(\Phi(t_i)) \rightarrow$$
$$\rightarrow \int_0^1 g(\Phi(t))dt = \int_{I_d} g(x)\, dx, \tag{8}$$

since $\{t_i\}_{i=1}^n$ are EQD, $g(\Phi(.))$ is also continuous, while last equality follows from SFC 1). The proof of WD property uses the Weyl criterion, generalized to WD that is too complicated to be presented here (see [17] for details).

Corollary 1.
Under the same assumptions as in Theorem 1, for points generated by Algorithm 1 we have:

$$D_n \to 0 \quad \text{and} \quad D_n^* \to 0 \quad \text{as} \quad n \to \infty \tag{9}$$

The first statement follows from (8) by selecting g as indicator functions of parallelepipeds. The second one is a direct consequence of (4).

Points generated by Algorithm 1, using the Hilbert SFC and $\theta = (\sqrt{5}-1)/2$ are shown in Fig. 2. We do not have a closed form formula for D_n^* of sequences generated by Algorithm 1. Instead, we have performed extensive simulations for n ranging from 50 to 30 000. Then, values of D_n^* were calculated and the least-squares method was used to fit the dependence of $D_n^*(d)$ on n for $d = 2$ and $d = 3$. The results are the following:

$$0.2\ \log^2(n)/n \quad \text{for} \quad d = 2, \tag{10}$$

$$0.06\ \log^3(n)/n \quad \text{for} \quad d = 3. \tag{11}$$

They follow a general pattern of "good" EQD sequences

$$D_n^*(d) = O\left(\frac{\log^d(n)}{n}\right). \tag{12}$$

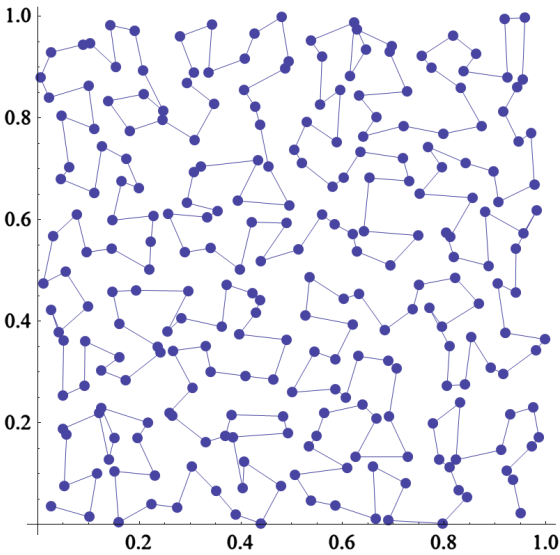


Fig. 2. $n = 256$ points generated by Algorithm 1, using the Hilbert SFC, and linked by thin lines, indicating their ordering on SFC that can be used as ordering for transmitting information between sensors.

5 Nonparametric Estimation of Spatial Means of Random Fields

Our aim in this section is to show how to estimate the spatial mean of a random field using WD points generated by Algorithm 1. The proposed estimator is similar to those proposed earlier in [12,13], but it differs in the following two respects.

- In [12] and [13] points were generated using the Halton-Hammersley sequences, which are EQD, but it is not known whether they are WD or not, while our sensors' positions are WD.
- Here, we allow correlated observations.

Let us assume that observations y_i of a scalar random field with unknown mean $f(x)$ are collected at spatial points x_i's , generated by Algorithm 1. More precisely,

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{13}$$

where ε_i are random variables (r.v.'s) for which the following conditions hold:

ERR1) ε_i's have zero mean, finite variance σ^2,
ERR2) for each $i = 1, 2, \ldots, n$ the covariance $E(\varepsilon_i \varepsilon_j)$ is not equal to zero only for a finite number, $0 \leq G(n) < n$ of indices $j = 1, 2, \ldots, n$.

This assumption means that we allow for correlations between observations from i-th sensor and observations from only a finite number $G(n)$ of its nearest neighborhood sensors with indices in a set denoted by $\mathcal{O}(i)$. As we shall see later, we may allow $G(n)$ to grow with n, but rather slowly. It is a reasonable assumption since the statistical dependence between sensors fades out rapidly as a function of a distance between them. Clearly, if errors are uncorrelated, we have $\mathcal{O}(i) = \{i\}$.

Our aim is to estimate f from (x_i, y_i), $i = 1, 2, \ldots, n$ in a nonparametric way, i.e., without imposing any finite parametrization of f. Instead, we impose some smoothness assumptions on f and our aim is to construct estimator $\hat{f}_n(x)$ such that the integrated mean-square error (IMSE) $I(f, f_n) \stackrel{def}{=} \int_{I_d} E(f(x) - \hat{f}_n(x))^2 \, dx \; \to \infty$ as $n \to \infty$.

We select the following estimator $\hat{f}_n$ of f.

Algorithm 2

$$\hat{f}_n(x) = \sum_{k=1}^{N} \hat{a}_{kn} v_k(x), \tag{14}$$

$$\hat{a}_{kn} = \frac{1}{n} \sum_{i=1}^{n} y_i v_k(x_i), \quad k = 1, 2, \ldots, N, \tag{15}$$

where $v_1, v_2, \ldots$ is a complete sequence of orthonormal functions[1] in the space $L_2(I_d)$ of square integrable functions on I_d. Denote by $a_k = \int_{I_d} f(x) v_k(x) \, dx$,

[1] Orthonormality means that $\int v_k(x)^2 \, dx = 1$ and $\int v_k(x) v_\ell(x) \, dx = 0$, for all $k \neq \ell$

$k = 1, 2, \ldots$, the coefficients of the Fourier series of f in the basis (v_k), i.e.,

$$f(x) \sim \sum_{k=1}^{\infty} a_k \, v_k(x), \tag{16}$$

where convergence is understand in L_2 norm. In (14) one can recognize the truncated version of (16). In practice, the truncation point N is also estimated. In asymptotical considerations below, we admit a slowly growing dependence of N on the number of sensors n and we shall write $N(n)$.

Remark 1. Notice that $\hat{a}_{kn}$ is the estimator of a_k, which is asymptotically unbiased. Indeed, equidistribution of x_i's implies

$$E(\hat{a}_{kn}) = n^{-1} \sum_{i=1}^{n} f(x_i) \, v_k(x_i) \to \int_{I_d} f \, v_k = a_k. \tag{17}$$

But, due tu WD property of x_i's, we can say more, namely

$$E(\hat{a}_{kn}) = n^{-1} \sum_{i=p}^{n+p} f(x_i) \, v_k(x_i) \to a_k, \tag{18}$$

uniformly w.r.t p. This property seems to be important for WSN, because it frequently happens that a group of sensors fails (due to, e.g., a battery or communication faults) and can be replaced by differently located sensors. From (18) it follows that WSN based on sensors placed at WD points still can provide estimators with a small bias, provided that n, i.e., the number of active sensors, is sufficiently large.

Remark 2. Notice that (15) has the form that is well suited for collecting data along a network, because coefficients can be calculated recursively. Indeed, $v_k(x_i)$' s, $(k = 1, 2, \ldots, N)$ can be pre-computed and stored in i-h sensor. Sensors can be ordered along SFC (see Fig. 2) and values of partial sums, after adding $y_i \, v_k(x_i)$ to previous ones can be passed from one sensor to another. The role of the sink is to calculate (14).

Now, our aim is to sketch the proof of IMSE consistency of $\hat{f}_n$. For simplicity of formulas, we assume that $d = 2$ and ONS) $v_k(x^{(1)}, x^{(2)})$ are ordered and normalized trigonometric functions of the form: 1, $\sin(x^{(1)})$ $\sin(x^{(2)})$, $\sin(x^{(1)})$ $\cos(x^{(2)}), \cos(x^{(1)})$ $\sin(x^{(2)})$, $\cos(x^{(1)})$ $\cos(x^{(2)}), \ldots$. We have $N = N(n)$ such products in common and they are commonly bounded by $H > 0$, say, while their derivatives are commonly bounded by $N(n) \, H$. Notice that $N(n)$ implicitly depends on d, so for $d = 2$, $N(n) = M_1(n) \, M_2(n)$, where $M_1(n)$ and $M_2(n)$ are the numbers of trigonometric functions in $x^{(1)}$ and $x^{(2)}$ variables, respectively. Below, when we write $N(n) \to \infty$, we require that $M_1(n) \to \infty$ and $M_2(n) \to \infty$.

One can weaken this assumption by allowing H be dependent on k, as it is, e. g., for the Legendre polynomials.

The orthonormality of (v_k) implies

$$IMSE(\hat{f}_n, f) = W_n + B_n^2 + R(N, f) \tag{19}$$

where $R(N,f) \stackrel{def}{=} \sum_{k=N+1}^{\infty} a_k^2$,

$$W_n \stackrel{def}{=} \sum_{k=1}^{N} \mathrm{Var}(\hat{a}_{kn}), \tag{20}$$

$$B_n^2 \stackrel{def}{=} \sum_{k=1}^{N} (E(\hat{a}_{kn}) - a_k)^2, \tag{21}$$

To estimate the variance, let us note that

$$\mathrm{Var}(\hat{a}_{kn}) = \frac{1}{n^2} E\left(\sum_{i=1}^{n} \varepsilon_i \, v_k(x_i)\right)^2 = \tag{22}$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} cov(\varepsilon_i, \varepsilon_j) \; v_k(x_i) \; v_k(x_j) =$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \in \mathcal{O}(i)} cov(\varepsilon_i, \varepsilon_j) \; v_k(x_i) \; v_k(x_j)$$

Due to ONS) and ERR2) we have

$$\mathrm{Var}(\hat{a}_{kn}) \leq \frac{H \, G(n) \, \sigma^2}{n} \left[\frac{1}{n} \sum_{i=1}^{n} |v_k(x_i)|\right] \leq \frac{H^2 \, G(n) \, \sigma^2}{n}. \tag{23}$$

Thus, $W_n \leq \frac{N(n) \, H^2 \, G(n) \, \sigma^2}{n}$.

Denote by $V(I_d)$ the space of functions having bounded variation and let $\mathcal{V}(g)$ denote the total variation of $g \in V(I_d)$. Then, by the Koksma-Hlavka inequality (see, e.g., [6]) we obtain for $f \in V(X) \cap C(X)$

$$(E(\hat{a}_{kn}) - a_k)^2 \leq (\mathcal{V}(f \cdot v_k) \cdot D_n^*)^2 \leq c_1 \, N^2(n) \, (D_n^*)^2, \tag{24}$$

where $c_1 > 0$ is a constant independent of n, while the second inequality can be obtained in a way similar to the one that was used in [12]. Thus, $B_n^2 \leq c_1 \, N^3(n) \, (D_n^*)^2$ and finally, we obtain

$$IMSE(\hat{f}_n, f) \leq c_1 \, N^3(n) \, (D_n^*)^2 + c_2 \, N(n)/n + R(N,f). \tag{25}$$

Notice that $f \in V(I_d) \cap C(I_d) \subset L^2(I_d)$, which implies $R(N,f) \to 0$ as $N(n) \to \infty$.

Proposition: If $f \in V(I_d) \cap C(I_d)$ and sequence $N(n)$ is selected in such a way that for $n \to \infty$

$$N(n) \to \infty, \quad N^3(n) \, (D_n^*)^2 \to 0, \quad N(n)/n \to 0, \tag{26}$$

then $IMSE(\hat{f}_n, f) \to 0$.

In the case of bivariate random fields, a particular choice of $N(n) = M_1(n)$ $M_2(n)$ depends on the rate of decay D_n^*, which is typically of order $O\left(\frac{1}{n^{1-\epsilon}}\right)$, where $\epsilon > 0$ is arbitrarily small. Thus, selecting $M_1(n) = M_2(n) = c\, n^{\beta/2}$, $c > 0$, $0 < \beta < 2\,(1-\epsilon)/3$, we can assure that for $N(n) = M_1(n)\, M_2(n)$ conditions (26) hold.

References

1. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: a survey. Comput. Netw. **38**, 393–422 (2002)
2. Butz, A.R.: Alternative algorithm for Hilbert's space-filling curve. IEEE Trans. Comput. **C–20**, 424–426 (1971)
3. Chong, C.-Y., Kumar, S.P.: Sensor networks: evolution, opportunities, and challenges. Proc. IEEE **91**, 1247–1256 (2003)
4. Drmota, M., Tichy, R.F.: Sequences, Discrepancies and Applications. Springer, Heidelberg (1997)
5. Griffith D.A.: Statistical efficiency of model-informed geographic sampling designs. In: Caetano, M., Painho, M. (eds.) 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, pp. 91–98 (2006)
6. Kuipers, L., Niederreiter, H.: Uniform Distribution of Sequences. Wiley, New York (1974). (Reprint, Dover, Mineola (2006))
7. Lemieux, C.: Monte Carlo and Quasi-Monte Carlo Sampling. Springer, New York (2009)
8. Niederreiter, H.: Random Number Generation and QuasiMonte Carlo Methods. SIAM, Philadelphia (1992)
9. Patan, M.: Optimal sensor network scheduling in identification of distributed parameter systems. Lecture Notes in Control and Information Sciences. Springer, Heidelberg (2012)
10. Pilz, J., Spöck, G.: Spatial sampling design for prediction taking account of uncertain covariance structure. In: Caetano, M., Painho, M. (eds.) 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, pp. 109–119 (2006)
11. Rafajłowicz, E., Rafajłowicz, W.: Sensors allocation for estimating scalar fields by wireless sensor networks. In: Proceedings 2010 Fifth International Conference on Broadband and Biomedical Communications IB2Com, Malaga (2010)
12. Rafajłowicz, E., Schwabe, R.: Halton and Hammersley sequences in multivariate nonparametric regression. Stat. Probab. Lett. **76**, 803–812 (2006)
13. Rafajłowicz, E., Schwabe, R.: Equidistributed designs in nonparametric regression. Statistica Sinica **13**, 129–142 (2003)
14. Rasch, D., Pilz, J., Verdooren, L.R., Gebhardt, A.: Optimal Experimental Design with R. Francis and Taylor, Boca Raton (2011)
15. Sagan, H.: Space-Filling Curves. Springer, New York (1994)
16. Skubalska-Rafajłowicz, E.: Pattern recognition algorithm based on space-filling curves and orthogonal expansion. IEEE Trans. Inf. Theory **47**, 1915–1927 (2001)
17. Skubalska-Rafajłowicz, E., Rafajłowicz, E.: Sampling multidimensional signals by a new class of quasi-random sequences. Multidimension. Syst. Signal Process. **23**, 237–253 (2012)

18. Uciński, D.: Optimal Measurement Methods for Distributed Parameter System Identification. CRC Press, London (2005)
19. Uciński D.: An algorithm to configure a large-scale monitoring network for parameter estimation of distributed systems. In: Proceedings of the European Control Conference 2007, Kos, Greece (2007)

On the Diversity Order of UW-OFDM

Heidi Steendam(✉)

DIGCOM Research Group, TELIN Department, Ghent University,
Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium
Heidi.Steendam@telin.ugent.be
http://telin.ugent.be/

Abstract. Unique word (UW) OFDM is a multicarrier technique that follows a different approach than standard multicarrier techniques like cyclic prefix (CP) OFDM. It has been reported that UW-OFDM outperforms CP-OFDM in the sense that it performs better in fading channels [1,2], and it has much lower out-of-band radiation [3] compared to CP-OFDM. However, a theoretical analysis of the error rate performance of UW-OFDM was never addressed in the literature. In this paper, we derive analytical expressions for the bit error rate for UW-OFDM, from which we can obtain the diversity order. It turns out that when the code generator matrix, needed to construct the UW-OFDM signal, is full rank, the UW-OFDM system reaches the maximum diversity order. This is in contrast with standard CP-OFDM, where only diversity order one can be reached, unless additional precoding is applied. Further, in the paper, we propose a construction method for the code generator matrix to achieve a (close to) maximum coding gain.

1 System Description

In multicarrier techniques, typically a guard interval is used to avoid intersymbol interference between successively transmitted symbols. In standard multicarrier techniques such as CP-OFDM, this guard interval is added on top of the DFT interval, implying the length of a transmitted symbol is increased. A different approach is used in UW-OFDM: here the guard interval is part of the DFT block. Further, in contrast with CP-OFDM, where the guard interval samples depend on the transmitted data, and are thus a priori unknown to the receiver, the guard interval in UW-OFDM is filled with known samples. The UW-OFDM signal is constructed in two steps. First, the data is modulated on the carriers such that after the N-point DFT, the last N_u time domain samples are zero. In this zero part of the signal, the unique word consisting of N_u known samples will be added. In order to obtain the zeroes in the time domain, we have to introduce redundancy in the frequency domain. Assume that because of the presence of

The author gratefully acknowledges the financial support from the Flemish Fund for Scientific Research (FWO). This research has been funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office.

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 317–326, 2014.
DOI: 10.1007/978-3-662-45504-3_31

guard bands, $N_m \leq N$ carriers are modulated. In that case maximally $N_m - N_u$ data symbols can be transmitted per DFT interval.

Let us assume we transmit $N_d \leq N_m - N_u$ data symbols $\mathbf{x}_d$. The required redundancy in the frequency domain is added by multiplying the data symbols with the $N_m \times N_d$ code generator matrix $\mathbf{G}$. To select which carriers are modulated and which are reserved as guard carriers, we use the $N \times N_m$ carrier selection matrix $\mathbf{B}$, where $N_m \leq N$ is the number of modulated carriers. This carrier selection matrix is a reduced version of the $N \times N$ identity matrix, where the columns corresponding to the unmodulated carriers are deleted. The resulting frequency domain vector is applied to the inverse DFT, resulting in the time domain samples

$$\mathbf{y} = \mathbf{F}_N^H \mathbf{B}\mathbf{G}\mathbf{x}_d = \begin{pmatrix} * \\ \mathbf{0} \end{pmatrix} \tag{1}$$

where $(\mathbf{F}_N)_{k,\ell} = \frac{1}{\sqrt{N}} e^{-j2\pi \frac{k\ell}{N}}$. The last N_u elements of $\mathbf{y}$ must be zero, implying that the matrix $\mathbf{G}$ must belong to the null space of the matrix $\tilde{\mathbf{F}}$, which consists of the N_u bottom rows of $\mathbf{F}_N^H \mathbf{B}$. As $\mathbf{F}_N^H$ is an orthogonal matrix and $\mathbf{B}$ is full rank, also the submatrix $\tilde{\mathbf{F}}$ has full rank, inferring the null space has dimension $N_m - N_u$. Let us define the $N \times (N_m - N_u)$ matrix $\mathbf{U}$ as the matrix containing an orthonormal basis for this null space. Such a basis can easily be found using the singular value decomposition of $\tilde{\mathbf{F}}$. As the columns of the matrix $\mathbf{G}$ belong to this null space, the matrix $\mathbf{G}$ can be written as the following linear combination:

$$\mathbf{G} = \mathbf{U}\mathbf{W} \tag{2}$$

where the $(N_m - N_u) \times N_d$ matrix $\mathbf{W}$ can freely be selected.

2 Theoretical Error Performance

In this section we derive an upper bound on the bit error rate when the UW-OFDM signal is transmitted over a Rayleigh fading channel. The channel is modelled as a tapped delay line with $L + 1$ taps: $\mathbf{h} = [h(0) \ldots h(L)]^T$, and the channel adds white Gaussian noise with spectral density $N_0/2$ per real dimension. To avoid intersymbol interference, we assume that the guard interval, i.e. the unique word, is longer than the channel length: $N_u \geq L$. Neglecting the presence of the unique word, the received sequence, is applied to a DFT resulting in the samples

$$\mathbf{r} = \mathbf{F}_N \mathbf{H} \mathbf{F}_N^H \mathbf{B}\mathbf{G}\mathbf{x}_d + \mathbf{F}_N \mathbf{w} = \tilde{\mathbf{H}}\mathbf{B}\mathbf{G}\mathbf{x}_d + \mathbf{F}_N \mathbf{w}. \tag{3}$$

where $\tilde{\mathbf{H}}_{k,k'} = \delta_{k,k'} \sum_{\ell=0}^{L} h(\ell) e^{j2\pi \frac{k\ell}{N}}$.

To derive the error rate performance, we use a similar approach as in [4]. Let us define the pairwise error probability (PEP) of the transmitted data vector $\mathbf{x}_d$ and the detected data vector $\mathbf{x}'_d \neq \mathbf{x}_d$, given the channel realization $\mathbf{h}$, by $Pr(\mathbf{x}'_d \neq \mathbf{x}_d | \mathbf{h})$. This PEP can be upper bounded using the Chernoff bound:

$$Pr(\mathbf{x}'_d \neq \mathbf{x}_d | \mathbf{h}) \leq \exp\left(-\frac{d^2(\mathbf{v}, \mathbf{v}')}{4N_0}\right) \tag{4}$$

with $d^2(\mathbf{v}, \mathbf{v}')$ the Euclidean distance between the vectors $\mathbf{v}$ and $\mathbf{v}'$ that depend on the transmitted and detected data sequences: $\mathbf{v} = \tilde{\mathbf{H}}\mathbf{B}\mathbf{G}\mathbf{x}_d$ and $\mathbf{v}' = \tilde{\mathbf{H}}\mathbf{B}\mathbf{G}\mathbf{x}'_d$. Defining $\tilde{\mathbf{H}}\mathbf{B}\mathbf{G}(\mathbf{x}_d - \mathbf{x}'_d) = \tilde{\mathbf{H}}\mathbf{B}\mathbf{G}\mathbf{e} = \mathbf{B}_e\mathbf{h}$, we rewrite the Euclidean distance as $d^2(\mathbf{v}, \mathbf{v}') = \mathbf{h}^H\mathbf{B}_e^H\mathbf{B}_e\mathbf{h}$. Averaging over the random channel, an upper bound on the average PEP is found:

$$Pr(\mathbf{x}_d \neq \mathbf{x}'_d) \leq \prod_{\ell=0}^{L} \frac{1}{1 + \alpha_L \frac{\lambda_{e,\ell}}{4N_0}} \tag{5}$$

where $\lambda_{e,\ell}$ are the eigenvalues of the matrix $\mathbf{B}_e^H\mathbf{B}_e$. We assumed in the derivation of (5) that the channel taps were uncorrelated: $E[\mathbf{h}\mathbf{h}^H] = \alpha_L^2\mathbf{I}_{L+1}$, with $\alpha_L = \frac{1}{L+1}$. The upper bound on the average PEP still depends on the unknown error vector $\mathbf{e} = \mathbf{x}_d - \mathbf{x}'_d$ through the eigenvalues. We assume, without loss of generality, that $\mathbf{e}^H\mathbf{e} = 1$.

Because of the definition of $\mathbf{B}_e^H\mathbf{B}_e$, i.e., $d^2(\mathbf{v}, \mathbf{v}') = \mathbf{h}^H\mathbf{B}_e^H\mathbf{B}_e\mathbf{h} \geq 0$, it follows that the eigenvalues are real-valued and non-negative. Assuming there are r_e non-zero eigenvalues $\lambda_{e,\ell} > 0$, we can further upper bound (5):

$$Pr(\mathbf{x}_d \neq \mathbf{x}'_d) < \left(\frac{1}{4N_0}\right)^{-r_e} \left(\prod_{\ell=1}^{r_e} \alpha_L \lambda_{e,\ell}\right)^{-1}. \tag{6}$$

The first factor determines the diversity order, i.e., the diversity order equals r_e, and the second factor is related to the coding gain γ_e. If there are no zero eigenvalues, i.e. when $\mathbf{B}_e^H\mathbf{B}_e$ is full rank, the diversity order is maximized: $r_e = L+1$. In that case, the coding gain equals $\gamma_e = \alpha_L \left[\det(\mathbf{B}_e^H\mathbf{B}_e)\right]^{\frac{1}{L+1}}$. The coding gain and diversity order still depend on the unknown error vector $\mathbf{e}$. The maximum obtainable diversity order and coding gain, irrespective of the data sequence is obtained by minimizing r_e and γ_e over the data.

$$\begin{aligned} r &= \min_{\mathbf{e}\neq\mathbf{0}} r_e \\ \gamma &= \min_{\mathbf{e}\neq\mathbf{0}} \gamma_e. \end{aligned} \tag{7}$$

It has been shown in [4] that the maximum diversity order can be achieved provided that the minimum Euclidean distance is larger than the channel length: $d_{\min} = \min_{\mathbf{e}} d(\mathbf{v}, \mathbf{v}') \geq L+1$. It turns out that a sufficient condition to reach the maximum diversity order is that $\text{rank}(\mathbf{BG}) = N_d$, and $N_d \geq L+1$, implying that the matrix $N_m \times N_d$ matrix $\mathbf{G}$ must be full rank. Taking into account the decomposition $\mathbf{G} = \mathbf{UW}$, where the matrix $\mathbf{U}$ is full rank because it consists of orthogonal basis vectors of the null space, it follows that the matrix $\mathbf{W}$ must be full rank to obtain full diversity. Hence, it turns out to be quite simple to design a UW-OFDM system with full diversity: it can be shown that the standard implementations for UW-OFDM, given in [1] and [2], reach full diversity. This is in contrast with CP-OFDM, which reaches a diversity of one only, unless we apply precoding.

3 Coding Gain

In the following, we restrict our attention to the case where the code generator matrix is full rank, i.e. the system has full diversity. We are interested in the code generator matrix that maximizes the coding gain γ. Comparing the upper bounds (4) and (5), it is obvious that if we find a code generator matrix $\mathbf{G}$ that maximizes the minimum Euclidean distance $d^2(\mathbf{v}, \mathbf{v}')$ irrespective of the error vector, also the coding gain will be (close to) maximum. Hence, let us look closer at the Euclidean distance. Given that $\mathbf{v} = \tilde{\mathbf{H}}\mathbf{B}\mathbf{G}\mathbf{x}_d$ and $\mathbf{v}' = \tilde{\mathbf{H}}\mathbf{B}\mathbf{G}\mathbf{x}'_d$, it follows that

$$d^2(\mathbf{v}, \mathbf{v}') = \mathbf{e}^H \mathbf{G}^H \mathbf{B}^H \tilde{\mathbf{H}}^H \tilde{\mathbf{H}} \mathbf{B} \mathbf{G} \mathbf{e} = \mathbf{e}^H \mathbf{A}_R \mathbf{e} \tag{8}$$

where the matrix $\mathbf{A}_R$ is a positive semi-definite Hermitian $N_d \times N_d$ matrix. In the following, we assume that the error vector $\mathbf{e} \in \mathbb{C}^{N_d \times 1}$, with $\mathbf{e}^H \mathbf{e} = 1$. Using the property that the Rayleigh quotient $(\mathbf{e}^H \mathbf{A}_R \mathbf{e})/(\mathbf{e}^H \mathbf{e})$ is bounded by the minimum and maximum eigenvalue of the matrix $\mathbf{A}_R$ [5], it follows that the minimum of the product $\mathbf{e}^H \mathbf{A}_R \mathbf{e}$ corresponds to the minimum eigenvalue of $\mathbf{A}_R$. If we want the minimum Euclidean distance to be as large as possible, this implies that the minimum eigenvalue of $\mathbf{A}_R$ must be as large as possible.

To maximize the minimum eigenvalue, we use the property of the Gerschgorin circles [5]. For a Hermitian matrix $\mathbf{A}_R$, this property states that the real-valued eigenvalues λ_k are located in the intervals $1 - R_k \leq \lambda_k \leq 1 + R_k$, with

$$R_k = \sum_{\substack{\ell=1 \\ \ell \neq k}}^{N_d} |(\mathbf{A}_R)_{k,\ell}|. \tag{9}$$

In the following we assume that $(\mathbf{G}^H \mathbf{B}^H \tilde{\mathbf{H}}^H \tilde{\mathbf{H}} \mathbf{B} \mathbf{G})_{k,k} = 1$. The normalization of the received energy per symbol implies that all data symbols have the same error performance [6,7], which results in the lowest error rate performance if we average over all data symbols. As a consequence, the sum of the eigenvalues λ_k is a constant: $\text{trace}(\mathbf{A}_R) = \sum_{k=1}^{N_d} \lambda_k = N_d$. It is straightforward to show through Lagrange optimization that maximizing the minimum eigenvalue under the constraint that the sum of the eigenvalues is known, corresponds to the case where all eigenvalues are equal, which infers $\mathbf{A}_R$ must be the identity matrix. This implies that we have to select the code generator matrix $\mathbf{G}$ such that

$$\mathbf{A}_R = \mathbf{G}^H \mathbf{B}^H \tilde{\mathbf{H}} \tilde{\mathbf{H}} \mathbf{B} \mathbf{G} = \mathbf{I}_{N_d}. \tag{10}$$

However, the matrix $\mathbf{A}_R$ depends on the channel taps $\mathbf{h}$ through the diagonal matrix $\tilde{\mathbf{H}}$. Hence, unless the channel is known, finding the code generator matrix that results in $\mathbf{A}_R = \mathbf{I}_{N_d}$ is generally not possible. Therefore, we restrict our attention to the case where we know the channel. This case could correspond to the case of a fixed wired link, or a wireless link with slowly varying channel, where the channel is estimated and fed back to the transmitter.

In the following, we propose a systematic construction method for the matrix $\mathbf{G}$ based on the decomposition $\mathbf{G} = \mathbf{U}\mathbf{W}$ (2). Note that the $N_m \times (N - N_u)$

matrix $\mathbf{U}$ is composed using the orthonormal basis vectors of the null space of the matrix $\tilde{\mathbf{F}}$, and forces the last N_u time domain samples in the DFT block to be zero. As soon as the system parameters (N, N_m and N_u) are known, the matrices $\mathbf{U}$ and $\mathbf{B}$ are fixed. Hence, we only need to select the matrix $\mathbf{W}$, which needs to be full rank to have full diversity. Using $\mathbf{G} = \mathbf{UW}$ in (10), we obtain the following restriction on $\mathbf{W}$: $\mathbf{W}^H\mathbf{U}^H\mathbf{B}^H\tilde{\mathbf{H}}^H\tilde{\mathbf{H}}\mathbf{BUW} = \mathbf{I}_{N_d}$. Let us consider the eigenvalue decomposition of the known $(N_m - N_u) \times (N_m - N_u)$ matrix $\mathbf{U}^H\mathbf{B}^H\tilde{\mathbf{H}}^H\tilde{\mathbf{H}}\mathbf{BU} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^H$. Because of the Hermitian nature of the matrix, its eigenvalues are real-valued and its eigenvector matrix $\mathbf{V}$ is a unitary matrix. Further, $\mathbf{U}^H\mathbf{B}^H\tilde{\mathbf{H}}^H\tilde{\mathbf{H}}\mathbf{BU}$ is the Gram matrix of $\tilde{\mathbf{H}}\mathbf{BU}$ implying the matrix is positive semi-definite. Assuming the diagonal matrix $\mathbf{B}^H\tilde{\mathbf{H}}^H\tilde{\mathbf{H}}\mathbf{B}$ is full rank, i.e., the channel at the frequencies of the modulated carriers does not contain spectral nulls, its eigenvalues are strictly non-zero. Without loss of generality, we can decompose $\mathbf{W}$ as $\mathbf{W} = \mathbf{VZ}$, resulting in $\mathbf{W}^H\mathbf{U}^H\mathbf{B}^H\tilde{\mathbf{H}}^H\tilde{\mathbf{H}}\mathbf{BUW} = \mathbf{Z}^H\mathbf{V}^H\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^H\mathbf{VZ} = \mathbf{Z}^H\boldsymbol{\Lambda}\mathbf{Z} = \mathbf{I}_{N_d}$. Further, defining $\boldsymbol{\Lambda} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}$, where the real-valued diagonal matrix $\boldsymbol{\Gamma}$ equals $\boldsymbol{\Gamma} = \text{diag}(\sqrt{\lambda_k})$, with λ_k the eigenvalues contained in $\boldsymbol{\Lambda}$, we can substitute $\mathbf{Z} = \boldsymbol{\Gamma}^{-1}\mathbf{X}$, resulting in:

$$\mathbf{X}^H\mathbf{X} = \mathbf{I}_{N_d}. \tag{11}$$

In the case of $N_r = N_u$, the matrix $\mathbf{X}$ is a square matrix. It can be verified that in this case the condition (11) can only be fulfilled when $\mathbf{X}$ is a unitary matrix. On the other hand, if $N_r > N_u$, the condition (11) requires that $\mathbf{X}$ is a finite frame [8]. In the following, we restrict our attention to the case where $N_u = N_r$.

4 Transmit Versus Received Power

In the previous section, we have proposed a systematic construction method to generate the code generation matrix that achieves (a close to maximum) coding gain if the channel is known at the transmitter. The degree of freedom that follows from the construction method (i.e., a unitary matrix $\mathbf{X}$ must be selected) gives us a large number of possible code generator matrices. In our solution, we have set a restriction on the received power, but no specifications were given on the transmit power. Let us look closer at the power at the transmitter P_T and the received power P_R:

$$\begin{aligned} P_R &= E_s\text{trace}(\mathbf{G}^H\mathbf{B}^H\tilde{\mathbf{H}}^H\tilde{\mathbf{H}}\mathbf{BG}) = N_dE_s \\ P_T &= E_s\text{trace}(\mathbf{G}^H\mathbf{B}^H\mathbf{BG}) \\ &= E_s\text{trace}(\mathbf{X}^H\boldsymbol{\Gamma}^{-1}\mathbf{V}^H\mathbf{U}^H\mathbf{B}^H\mathbf{BUV}\boldsymbol{\Gamma}^{-1}\mathbf{X}) \\ &= E_s\text{trace}(\mathbf{U}^H\mathbf{B}^H\mathbf{BUV}\boldsymbol{\Gamma}^{-1}\mathbf{XX}^H\boldsymbol{\Gamma}^{-1}\mathbf{V}^H) \\ &= E_s\text{trace}(\mathbf{U}^H\mathbf{B}^H\mathbf{BU}(\mathbf{U}^H\mathbf{B}^H\tilde{\mathbf{H}}^H\tilde{\mathbf{H}}\mathbf{BU})^{-1}). \end{aligned} \tag{12}$$

where in the last line, we used the unitary nature of the matrix $\mathbf{X}$, i.e., $\mathbf{XX}^H = \mathbf{I}_{N_d}$, and $\mathbf{V}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}^{-1}\mathbf{V}^H = \mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^H = (\mathbf{U}^H\mathbf{B}^H\tilde{\mathbf{H}}^H\tilde{\mathbf{H}}\mathbf{BU})^{-1}$. Note that the transmit power nor the received power depend on the selected unitary matrix $\mathbf{X}$, but only depend on the channel and system parameters, i.e., all

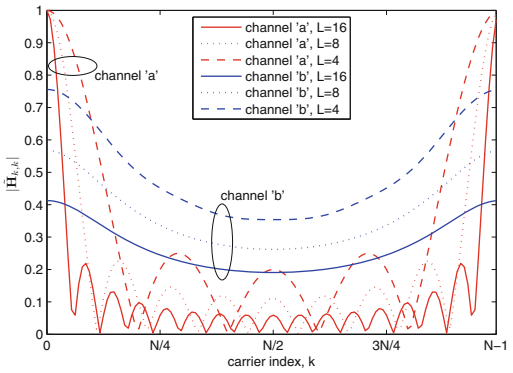


Fig. 1. Frequency response $|\tilde{\mathbf{H}}_{k,k}|$ of the channel for a) $h(\ell) = \nu$ and b) $h(\ell) = \nu \exp(-\ell)$; $\ell = 0, \ldots, L$, ν is a constant to normalize the channel: $\mathbf{h}^H\mathbf{h} = \alpha_L$.

matrices $\mathbf{X}$ lead to the same transmit/received power. In general, it turns out that the transmitted energy per symbol is not normalized. However, because $\mathbf{BG}$ needs to be full rank in order to have full diversity and as we assumed in our construction method that $\tilde{\mathbf{H}}$ does not contain spectral nulls at the modulated carriers, we know that in this case the transmit power is finite.

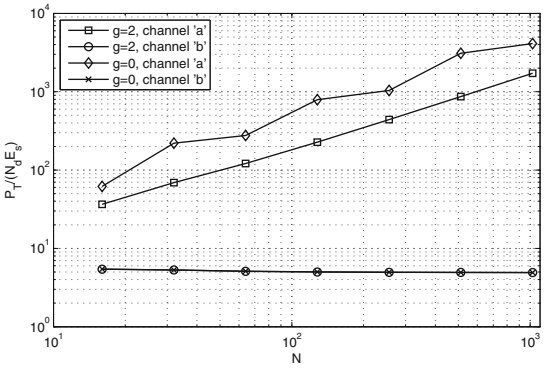
As an illustration, we evaluate the influence of the system parameters and the channel on the transmit power P_T for the following two channels:

$$\text{channel a} : h(\ell) = \nu$$
$$\text{channel b} : h(\ell) = \nu e^{-\ell}$$

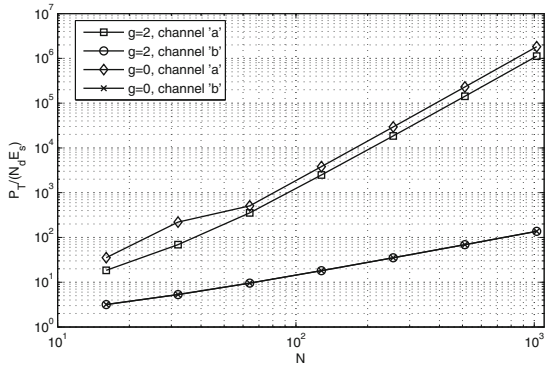
where the channel parameter ν is selected such that $\mathbf{h}^H\mathbf{h} = \frac{1}{L+1}$. Figure 1 shows the frequency response $|\tilde{\mathbf{H}}_{k,k}|$ for these two channels, for different values for L. In this figure, we observe that the frequency response of channel 'b' is reasonably flat, whereas channel 'a' is more frequency selective. None of the channels show spectral nulls[1], implying their channel matrix $\tilde{\mathbf{H}}$ is full rank. Further, it follows from the figure that the amplitude of the frequency response reduces when L increases.

Figure 2 shows the required transmit power normalized to the received power, i.e., $P_T/(N_d E_s)$, as function of the DFT size N, with and without guard band. The guard band consists of g unmodulated edge carriers at both sides of the frequency band. Hence, the number of modulated carriers equals $N_m = N - 2g$. In Fig. 2(a), the channel length is kept constant, while in Fig. 2(b), the channel length increases proportional to the DFT size. The transmit power for channel

[1] Although channel 'a' from Fig. 1 shows small values for the frequency response $|\tilde{\mathbf{H}}_{k,k}|$, they are non-zero. If we plot the y-axis in the figure with a logarithmic scale, it can be observed that the largest and smallest value of $|\tilde{\mathbf{H}}_{k,k}|$ differ approximately a factor of 100. This difference is not large enough to make the matrix $\tilde{\mathbf{H}}$ singular.



(a) $N_r = N_u = L = 4$



(b) $N_r = N_u = L = N/8$

Fig. 2. Transmit power P_T, normalized to the received power $N_d E_s$.

'a' is larger than for channel 'b', and the transmit power is larger for $g = 2$ than for $g = 0$, although the difference for channel 'b' is small. To explain this effect, look at the eigenvalues of the matrix $\mathbf{U}^H\mathbf{B}^H\tilde{\mathbf{H}}^H\tilde{\mathbf{H}}\mathbf{B}\mathbf{U}$ for both channels, with and without guard band, for $N = 32$. For other values of N, similar results are obtained. From Fig. 3, it follows that for both channels, the eigenvalues are larger for $g = 0$ than for $g = 2$, and the difference is larger in channel 'a' than in channel 'b'. This can be explained using Fig. 1. In our channel models, the amplitude of the channel frequency response is larger at the band edges. By not using these carriers, the average channel frequency response decreases, resulting in smaller eigenvalues. This effect is larger in channel 'a' because the channel is more frequency selective than channel 'b'. Further, the eigenvalues for channel 'a' are smaller than for channel 'b', as the frequency response of channel 'a' is for (almost) all carriers much smaller than for channel 'b'. Hence, it is expected that, to obtain the same received power, the required transmit power in channel 'a' will be larger than in channel 'b', and increasing the guard band width will result in an increased transmit power, which is confirmed in Fig. 2.

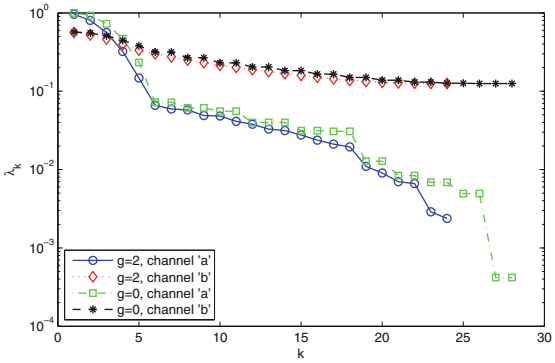


Fig. 3. Eigenvalues of the matrix $\mathbf{U}^H\mathbf{B}^H\tilde{\mathbf{H}}^H\tilde{\mathbf{H}}\mathbf{B}\mathbf{U}$ in decreasing order of magnitude, $N = 32$, $N_r = N_u = L = 4$

Further, in Fig. 2, it also can be observed that when the channel length is kept constant, the required transmit power for channel 'b' is essentially constant as function of the DFT size, whereas for channel 'a' the transmit power linearly increases with the DFT size. When the channel length increases with the DFT size, both the transmit power in channel 'a' and channel 'b' increase, although in channel 'a', the transmit power increases faster. To explain this effect, we consider the special case where all carriers are modulated, i.e., with $g = 0$. In that case, the carrier selection matrix $\mathbf{B}$ reduces to the identity matrix. In addition, the null space matrix $\mathbf{U}$ reduces to $\hat{\mathbf{F}}$, where $\hat{\mathbf{F}}$ corresponds to the first $N - N_u$ rows of $\mathbf{F}_N^H$. Consequently, the transmit power reduces to

$$P_T = E_s \text{trace}(\mathbf{U}^H\tilde{\mathbf{H}}^H\tilde{\mathbf{H}}\mathbf{U})^{-1}. \tag{13}$$

After some straightforward computations, it follows that for the two channels 'a' and 'b', the matrix $\mathbf{U}^H\tilde{\mathbf{H}}^H\tilde{\mathbf{H}}\mathbf{U}$ reduces to a symmetric banded Toeplitz matrix with elements $(\mathbf{U}^H\tilde{\mathbf{H}}^H\tilde{\mathbf{H}}\mathbf{U})_{m,m'} = \frac{1}{L+1}w(m-m')$, with

$$w_a(m) = \begin{cases} 1 - \frac{|m|}{L+1} & |m| \leq L \\ 0 & \text{else} \end{cases} \tag{14}$$

$$w_b(m) = \begin{cases} e^{-|m|}\frac{1-e^{-2(L+1-|m|)}}{1-e^{-2(L+1)}} & |m| \leq L \\ 0 & \text{else} \end{cases} \tag{15}$$

for channel 'a' and 'b', respectively[2]. For channel 'b', it follows from (15) that the matrix $\mathbf{U}^H\tilde{\mathbf{H}}^H\tilde{\mathbf{H}}\mathbf{U}$ is diagonally dominant. Hence, taking into account the Gerschgorin theorem [5], the $N - N_u$ eigenvalues λ_k of the matrix will be of

[2] For general channels, it can be verified that the matrix $\mathbf{U}^H\tilde{\mathbf{H}}^H\tilde{\mathbf{H}}\mathbf{U}$ is a banded matrix having L non-zero side diagonals at both sides of the main diagonal, although in general, the matrix is not symmetric Toeplitz.

the order of the diagonal elements, i.e., $\lambda_k \approx \frac{1}{L+1}$. This can also be observed in Fig. 3. The normalized transmit power $\frac{P_T}{N_d E_s} = \frac{1}{N_d}\text{trace}(\mathbf{U}^H\tilde{\mathbf{H}}^H\tilde{\mathbf{H}}\mathbf{U})^{-1} = \sum_{k=0}^{N-N_u-1}\frac{1}{\lambda_k}$ can therefore be approximated by $\frac{P_T}{N_d E_s} \approx L+1$, with $N_d = N-N_u$. For channel 'a', the matrix $\mathbf{U}^H\tilde{\mathbf{H}}^H\tilde{\mathbf{H}}\mathbf{U}$ is not diagonally dominant. However, in [9], it is shown that square banded Toeplitz matrices are asymptotically equivalent to a circulant matrix for increasing matrix size. Hence, the eigenvalues for channel 'a' can be approximated by the eigenvalues of the asymptotically equivalent circulant matrix, which are obtained by computing its spectrum $f(z)$:

$$\begin{aligned} f(z) &= \sum_{m=-L}^{L} \frac{1}{L+1} w_a(m) e^{jmz} \\ &= \left(\frac{\sin(L+1)\frac{z}{2}}{(L+1)\sin\frac{z}{2}}\right)^2. \end{aligned} \tag{16}$$

The eigenvalues of the matrix are given by $\lambda_k = f(\frac{2\pi k}{N-N_u})$. Hence, the eigenvalue spread will be larger than for channel 'b', which is also observed in Fig. 3. The transmit power is mainly determined by the smallest eigenvalues. These smallest eigenvalues correspond to the values of k for which $\frac{2\pi k}{N-N_u}$ is close to a zero of the function $f(z)$ in the interval $[0, 1[$, where the zeros are given by $z_m = \frac{2\pi m}{L+1}$, $m = 1, \ldots, \lfloor\frac{L+1}{2\pi}\rfloor$. To find the smallest eigenvalues, we derive the Taylor series expansion of $f(z)$ at $z = z_m$:

$$f(z) \approx \frac{(z-z_m)^2}{2\sin^2(\frac{\pi m}{L+1})}. \tag{17}$$

After some straightforward computations, we find that the smallest eigenvalues can be upper bounded by $\lambda_{\hat{k}} \leq 2[(N-N_u)\sin\frac{\pi m}{L+1}]^2$, where $\hat{k} = [\frac{\pi m(N-N_u)}{L+1}]_I$ and $[x]_I$ rounds x to the nearest integer. Taking this into account, the normalized transmit power for channel 'a' can be approximated by

$$\frac{P_T}{N_d E_s} \approx \frac{2(N-N_u)^2}{N_d} \sum_{m=1}^{\lfloor\frac{L+1}{2\pi}\rfloor} \sin^2\frac{\pi m}{L+1} \sim (N-N_u)(L+1). \tag{18}$$

Hence, the theoretical approximations for the transmit power for both channels confirm the behaviour of the transmit power in Fig. 2.

5 Conclusions

In this paper, we derived an analytical expression for the theoretical error rate performance for UW-OFDM. From this expression we obtained the conditions to achieve full diversity and a (close to) maximum coding gain: full diversity is reached if the code generator matrix is full rank, and a (close to) maximum coding gain requires that the matrix $\mathbf{A}_R$ must be the identity matrix. Based on

these conditions, we proposed a systematic construction method for the code generator matrix. We showed that the transmit power, given the received power is normalized, is independent of the selected code generator matrix, but only depends on the channel and the system parameters.

References

1. Huemer, M., Hofbauer, C., Huber, J.: Non-systematic complex number RS coded OFDM by unique word prefix. IEEE Trans. Sig. Process. **60**(1), 285–299 (2012)
2. Huemer, M., Onic, A., Hofbauer, C.: Classical and bayesian linear data estimators for unique word OFDM. IEEE Trans. Sig. Process. **59**, 6073–6085 (2011)
3. Rajabzadeh, M., Steendam, H., Khoshbin, H.: Power Spectrum Characterization of Systematic Coded UW-OFDM Systems. VTCFall2013, Las Vegas, 2–5 Sep 2013
4. Wang, Z., Giannakis, G.B.: Complex-field coding for OFDM over fading wireless channels. IEEE Trans. Inf. Theory **49**(3), 1–13 (2003)
5. Meyer, C.D.: Matrix Analysis and Applied Linear Algebra. SIAM, Philadelphia (2000)
6. Ngo, Q.-T., Berder, O., Scalart, P.: General minimum euclidean distance-based precoder for MIMO wireless systems. EURASIP J. Adv. Sig. Process. (2013). doi:10.1186/1687-6180-2013-39
7. Huang, X.-L., Wang, G., Hu, F.: Minimal euclidean distance-inspired optimal and suboptimal modulation schemes for vector OFDM system. Int. J. Commun. Syst. **24**, 553–567 (2011). doi:10.1002/dac.1171
8. Casazza, P.G., Leonhard, N.: Classes of finite equal norm parseval frames. Contemporary Math. **451**, 11–31 (2008)
9. Gray, R.M.: Toeplitz and Circulant Matrices: A review. Now Publishers Inc, Hanover (2013)

Representation and Analysis of Piecewise Linear Functions in Abs-Normal Form

Tom Streubel(✉), Andreas Griewank, Manuel Radons, and Jens-Uwe Bernt

Department of Mathematics, Humboldt University at Berlin, Berlin, Germany
{streubel,griewank,radons,berntj}@math.hu-berlin.de

Abstract. It follows from the well known min/max representation given by Scholtes in his recent Springer book, that all piecewise linear continuous functions $y = F(x) : \mathbb{R}^n \to \mathbb{R}^m$ can be written in a so-called abs-normal form. This means in particular, that all nonsmoothness is encapsulated in s absolute value functions that are applied to intermediate switching variables z_i for $i = 1, \ldots, s$. The relation between the vectors x, z, and y is described by four matrices Y, L, J, and Z, such that

$$\begin{bmatrix} z \\ y \end{bmatrix} = \begin{bmatrix} c \\ b \end{bmatrix} + \begin{bmatrix} Z & L \\ J & Y \end{bmatrix} \begin{bmatrix} x \\ |z| \end{bmatrix}$$

This form can be generated by ADOL-C or other automatic differentation tools. Here L is a strictly lower triangular matrix, and therefore z_i can be computed successively from previous results. We show that in the square case $n = m$ the system of equations $F(x) = 0$ can be rewritten in terms of the variable vector z as a linear complementarity problem (LCP). The transformation itself and the properties of the LCP depend on the Schur complement $S = L - ZJ^{-1}Y$.

Keywords: Piecewise linearization (PL) · Algorithmic differentiation (AD) · Equation solving · Semi-smooth newton · Smooth dominance · Complementary piecewise linear system (CLP) · Linear complementarity (LCP)

1 Introduction

Via algorithmic differentiation it is possible to calculate directional derivatives from *evaluation procedures* of vector valued functions simultaneously with their evaluation at a base point x_0. These evaluations are exact within the limitations of machine precision. An *evaluation procedure* is a composition of so called elementary functions, which are aggregated as a library in their symbolic form and thus make up the atomic constituents of complex functions. Basically the selection of elementary functions for the library is arbitrary, as long as they comply with assumption (ED) (elementary differentiability, in [3]), meaning that they are at least once Lipschitz-continuously differentiable. In the literature (see e.g. [3,8]) the following collection is suggested as the quasi-standard for a library:

$$\Phi = \{+, -, *, /, \sin, \cos, \tan, \cot, \exp, \log, \ldots\}$$

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 327–336, 2014.
DOI: 10.1007/978-3-662-45504-3_32

Common software packages such as ADOL-C provide tools for the algorithmic differentiation of functions composed from the contents of this collection.

But many practical problems and most algorithms are not smooth everywhere and thus cannot be modelled via a library that consists solely of a set of functions that comply with (ED). More specifically one is likely to encounter standard functions of computer arithmetic, that are not globally differentiable, e.g. abs, max and min. Since

$$\max(x, y, z) \equiv \max(\max(x, y), z), \quad \max(x, y) \equiv 0.5 * (x + y + \text{abs}(x - y))$$

max and min can be expressed in terms of the absolute value function. As shown, this reformulation of max and min provides us with a very practical handle on the representation of piecewise linearity, since Scholtes proved in [12], that any scalar-valued, real piecewise linear function $f : \mathbb{R}^n \to \mathbb{R}$ can be expressed as a finite nesting of max and min comparisons of linear functions. Here and throughout we use linear in the sense of affine, i.e. allow a constant increment.

Generally any one dimensional piecewise linear function $f : \mathbb{R} \to \mathbb{R}$ can be expressed in terms of absolute values. For a given set of points $\{(x_i, y_i) : i = 0, \ldots, n\}$, where $x_0 < x_1 < \cdots < x_n$, two outer slopes s_0, s_{n+1} and n inner slopes $s_i = (y_i - y_{i-1})/(x_i - x_{i-1})$, we obtain the formula

$$y = \frac{1}{2}\left[y_0 + s_0(x - x_0) + \sum_{i=0}^{n}(s_{i+1} - s_i)\,\text{abs}\,(x - x_i) + y_n + s_{n+1}(x - x_n)\right]$$

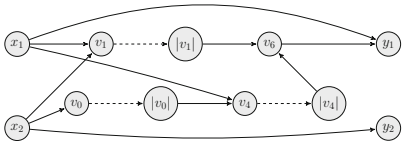
where the two linear functions at the beginning and the end can be combined to $[y_0 - s_0\,x_0 + y_n - s_{n+1}\,x_n + (s_0 + s_{n+1})x]/2$. This might be helpful for the purpose of implementation. For example with $a < b \in \mathbb{R}$ we obtain the cut-off function

$$f(x) = \max(a, \min(x, b)) = 0.5 * [a + \text{abs}(x - a) - \text{abs}(x - b) + b]$$

Similar to linear models of smooth functions, piecewise linearizations can be used to approximate piecewise smooth functions [12]. The aim is to extend the principles and techniques of classic algorithmic differentiation in such a way, that these piecewise linear models can be evaluated with the same efficiency, stability and simplicity of data structures as in the linear case. Since the absolute value function is already piecewise linear, it can be modelled by itself. By proposition 3.1 from [4] we have for the procedure (introduced in the next chapter) that the error of the piecewise linear approximation is of second order and varies Lipschitz continuously w.r.t. the developing point.

2 Piecewise Linearization and Abs-Normal Form

Example 1. Formula, graph and sequential code instruction of an *evaluation procedure*:



The v_i are called intermediate values. The indices are in a dependency relation $j \prec i$, if there is an edge from v_j to v_i. In general the values of a sequential code instruction of an *evaluation procedure* are denoted as a tuple

$$[v_{1-n}, v_{1-(n-1)}, \ldots, v_0, v_1, v_2, \ldots, v_i, \ldots, v_l] \quad \text{where}$$

$$\begin{aligned} v_{j-n} &= x_j && \text{for} \quad j = 1, \ldots, n \\ v_i &= \varphi_i(v_j)_{j \prec i} && \text{for} \quad i = 1, \ldots, l \quad \text{and} \quad \varphi \in \Phi_{\text{abs}} = \Phi \cup \{\text{abs}\} \end{aligned}$$

The values of the piecewise linearization can be evaluated simultaneuosly as increments of the function value by the following set of propagation rules [4] that implicitly defines a second code instruction.

Procedure 1.

$$[\Delta v_{1-n}, \Delta v_{1-(n-1)}, \ldots, \Delta v_0, \Delta v_1, \Delta v_2, \ldots, \Delta v_i, \ldots, \Delta v_l] \quad \textit{where}$$

$$\textit{for} \quad j = 1, \ldots, n: \quad \Delta v_{j-n} = \Delta x_j \quad \textit{and for} \quad i = 1, \ldots, l:$$

$$\begin{aligned} \Delta v_i &= \Delta v_j \pm \Delta v_k && \textit{when} && v_i = v_j \pm v_k \\ \Delta v_i &= \Delta v_j * v_k + v_j * \Delta v_k && \textit{when} && v_i = v_j * v_k \\ \Delta v_i &= c_{ij} * \Delta v_j && \textit{when} && v_i = \varphi(v_j) \end{aligned}$$

where $\varphi \in \Phi \setminus \{\pm, *, \text{abs}\}$ *and* $c_{i,j} = \varphi_i'(v_j)$ *is the local partial derivative*

$$\Delta v_i = \quad \text{abs}(v_j + \Delta v_j) - \text{abs}(v_j) \qquad \textit{when} \qquad v_i = \text{abs}(v_j)$$

Then the k-th component of the piecewise linearization is determined by:

$$y_k = v_{l-m+k} + \Delta v_{l-m+k}$$

The overall costs are at most four times of those of a function evaluation [3].

So far we have a method for a small number of evaluations at some basepoints. But for the purposes of integration, solving ODEs, optimization and solving piecewise linear equation systems (see [4,5]) we need a suitable data structure for a large number of evaluations of a single piecewise linearization. A general nonlinear concept of Barton and Khan from [6] combined with taping technology leads to the abs-normal form:

Definition 1. *For* $Z \in \mathbb{R}^{s \times n}, L \in \mathbb{R}^{s \times s}, J \in \mathbb{R}^{m \times n}, Y \in \mathbb{R}^{m \times s}$ *matrices, where* L *is of strictly lower triangular form and vectors* $c \in \mathbb{R}^s, b \in \mathbb{R}^m$*, the system*

$$\begin{bmatrix} z \\ y \end{bmatrix} = \begin{bmatrix} c \\ b \end{bmatrix} + \begin{bmatrix} Z & L \\ J & Y \end{bmatrix} \begin{bmatrix} x \\ |z| \end{bmatrix} \tag{1}$$

is called abs-normal form. The modulus operation $|z|$ has to be understood componentwise here. An abs-normal form is called simply switched if $L = 0$.

The components of z can be evaluated successively, since L is a strictly lower triangular matrix. The control flow in the evaluation of the abs-normal form is conveniently characterised by the signature vectors and matrices

$$\sigma_x \equiv \sigma_z \equiv \operatorname{sign}(z) \in \{-1, 0, 1\}^s, \qquad \Sigma_z = \operatorname{diag}(\sigma_z) \in \{-1, 0, 1\}^{s \times s}$$

In particular we will use throughout the identity $|z| = \Sigma_z z$. Using this relation we can eliminate z for any given $x \in \mathbb{R}^n$ and obtain the explicit representation

$$F(x) = y = \overbrace{b + Y\Sigma_\sigma (I - L\Sigma_\sigma)^{-1} c}^{\text{piecewise constant}} + J_\sigma \cdot x \tag{2}$$

$$\text{where } J_\sigma = J + Y\Sigma_\sigma (I - L\Sigma_\sigma)^{-1} Z \tag{3}$$

On the other hand every piecewise linear function in max-min expression can be represented in abs-normal form. Thus the abs-normal form is an equivalent characterization of piecewise linear mappings, which is stable w.r.t to perturbations[1]. Each signature vector $\sigma \in \{-1, 0, 1\}^s$ uniquely characterises the polyhedron

$$P_\sigma = \{x \in \mathbb{R}^n \mid \sigma_x = \sigma\}$$

The collection of these mutually disjoint and relatively open polyhedra forms a so called polyhedral decomposition or skeleton $\mathcal{P}$ of $\mathbb{R}^n$. The restriction of F to the closure of any $P_\sigma \in \mathcal{P}$ is linear (Fig. 1).

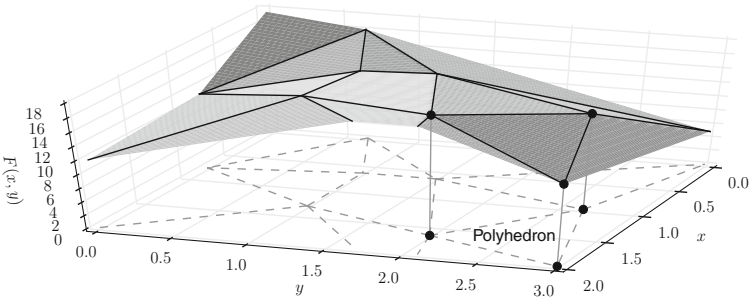


Fig. 1. Example of a piecewise linear function and its corresponding polyhedral decomposition

[1] Perturbations of the data Z, L, J, Y, c and b preserves the property of being a continuous, piecewise linear abs-normal form, provided L stays strictly lower triangular.

Each $P = P_\sigma$ has a nonempty interior if and only if it is open, in which case we will also refer to σ as open. By continuity all σ that have no zero components are open, but the converse need not be the case. It can be shown, that the J_σ given in (3) are limiting Jacobians in the following sense exactly if σ is open.

For general Lipschitz continuous F it follows from Rademacher's Theorem that it has a Frechet derivative $F'(x)$ at all points in a set D_F, whose complement has the measure zero. The set of limiting Jacobians at any $x_0 \in \mathbb{R}^n$ is defined as

$$\partial^L F(x_0) = \left\{ \lim_{\substack{x \to x_0 \\ x \in D_F}} F'(x) \right\} \neq \emptyset$$

and the set of generalized Jacobians in the sense of Clarke as

$$\partial F(x_0) = \mathrm{conv}(\partial^L F(x_0))$$

The definition of $\partial^L F(x_0)$ looks quite nonconstructive and in fact there is no general methodology for evaluating limiting Jacobians since the rules for propagating generalized derivatives are only inclusions. Given the abs-normal form one can compute limiting Jacobians that are also generalized Jacobians of the underlying nonlinear functions by a technique called polynomial escape [4,6]. The computational complexity is similar to that of the foward mode in the smooth case. Especially for generalized gradients where $m = 1$ an adaption of the much cheaper reverse mode is under development.

Throughout the remainder of this paper, we will only consider piecewise linear F in abs-normal form that are square in that $m = n$. Furthermore we assume w.l.o.g. that the so called smooth part J is nonsingular. If this is not a priori true one can shift terms by using the identity $x = \mathrm{abs}(x + \mathrm{abs}(x)) - \mathrm{abs}(x)$. The Schur complement of J within the abs-normal form is given by $S = L - ZJ^{-1}Y$. By using the Sherman-Morrison-Woodbury formula we can characterise the nonsingularity of the generalized Jacobian J_σ as follows

$$\det(J_\sigma) = \det(J) \det(I - S\Sigma_\sigma), \text{ for } \sigma = \sigma_x \in \{-1, 0, 1\}^s \tag{4}$$

Note that the upper half of the abs-normal form, which maps x onto z, need not be surjective. Hence the mapping is maybe partially switched in that some signature vectors $\sigma \in \{-1, 0, 1\}^s$ do not arise as σ_x for any x. In other words some P_σ might be empty. On the other hand if the linear map Zx is surjective, then the abs-normal form must be totally switched in that all 3^n sign combinations of σ with corresponding nonempty P_σ do arise. The following so called complementary piecewise linear mappings are always totally switched, since $z \in \mathbb{R}^s$ becomes independent and ranges over all of $\mathbb{R}^s$.

3 Complementary Piecewise Linear Systems and Their Relation to LCPs

In contrast the nonsingularity of the smooth part J allows the elimination of x for any given z and y.

$$y = b + Jx + Y|z| \iff x = J^{-1}(y - b) - J^{-1}Y|z|$$

In view of solving $F(x) = 0$ we can set $y = 0$ or absorb it into b. Then substitution of x into the upper half yields the complementary piecewise linear mapping

$$H(z) \equiv (I - S\Sigma_z)z - \hat{c}, \text{ where } \hat{c} \equiv c - ZJ^{-1}b$$

The function $H : \mathbb{R}^s \to \mathbb{R}^s$ is still piecewise linear and has the abs-normal form

$$\begin{bmatrix} \tilde{z} \\ H(z) \end{bmatrix} = \begin{bmatrix} 0 \\ -\hat{c} \end{bmatrix} + \begin{bmatrix} I & 0 \\ I & -S \end{bmatrix} \begin{bmatrix} z \\ |\tilde{z}| \end{bmatrix} \tag{5}$$

whose Schur complement is again S. Since the new L vanishes, the complementary piecewise linear map is always simply switched. Moreover the polyhedral decomposition consists entirely of 2^n open orthants and their faces. As shown in [5] this implies that H is bijective if and only if it is an open map. For general PL functions and in particular the underlying F we only have the chain of implications [12]

$$F \text{ is injective} \implies F \text{ is open} \implies F \text{ is surjective}$$

Furthermore Scholtes has proven in [12] that piecewise linear maps are open maps if and only if the determinants of all limiting Jacobians have the same sign (are w.l.o.g. positive). The limiting Jacobians of H are exactly the shifted identities $I - S\Sigma$ for any $\Sigma = \text{diag}(\sigma)$ with $\sigma \in \{-1, 1\}^s$. Consequently, coherent orientation of H occurs if and only if all $\det(I - S\Sigma)$ are positive, which implies by (4) the coherent orientation of F. Whereas the converse need not be true, i.e. F may be coherently oriented but H not.

The problem of solving $H(z) = 0$, for some $z \in \mathbb{R}^n$ can be recast as a linear complementarity problem (LCP). It turns out to have the P-matrix property if and only if H is coherently oriented [11]. The reformulation requires:

Lemma 1. *Let $M, S \in \mathbb{R}^{s\times s}$ arbitrary, s.t. $(I + S)M = (I - S)$, then*

1. $\det(I + S) \neq 0 \iff \det(I + M) \neq 0$
2. $S = (I + M)^{-1}(I - M)$ *if* $\det(I + M) \neq 0$

Proof.

$$\begin{aligned} M = [I + S]^{-1}[I - S] &\iff [I + S]\tfrac{1}{2}(I + M) = \tfrac{1}{2}([I + S] + [I - S]) = I \\ &\iff S = 2(I + M)^{-1} - I = (I + M)^{-1}(I - M) \end{aligned}$$ □

Now consider two vectors $0 \leq u, w \geq 0$, such that $z = u - w$ and $u^\top w = 0$. Then by the upper half of an abs-normal form for F

$$\begin{aligned} u - w &= c + Zx + L(u + w) \\ &= c + \left[ZJ^{-1}(y - b) - ZJ^{-1}Y(u + w)\right] + L(u + w) \\ &= \hat{c} + S(u + w) \iff (I - S)u = \hat{c} + (I + S)w \\ \iff \quad w &= u - (I + S)^{-1}\hat{c} \iff w = Mu + q \end{aligned}$$

where $M \equiv (I+S)^{-1}(I-S)$ and $q \equiv -(I+S)^{-1}\hat{c}$. Because of the substitution of x, the solutions of this standard LCP $w = q + Mu$, are solutions of the complementary piecewise linear system H. Any standard LCP $w = q + Mu$, where $u, w \geq 0$ and $u^\top w = 0$, can be rewritten as a complementary piecewise linear equation system as

$$z = (I+M)^{-1}(I-M)|z| - 2(I+M)^{-1}q$$

where $u = \frac{1}{2}(|z| + z)$ and $w = \frac{1}{2}(|z| - z)$. This was proven by Bokhoven in his thesis [1]. To transform the complementary piecewise linear system into an LCP or vice versa one has to compute the Möbius transform of S or M, respectively. This requires in either case at least implicitly a matrix inversion and several multiplications. Therefore we consider methods for directly solving the original and complementary piecewise linear system possibly even avoiding the explicit computation of $S = L - ZJ^{-1}Y$.

4 Solving Piecewise Linear Equation Systems

The principal task is to find solutions $x \in \mathbb{R}^n$, such that $F(x) = 0$ with piecewise linear $F : \mathbb{R}^n \to \mathbb{R}^n$. A possible nonzero right hand side can be absorbed into the vector b as described above.

There are several methods developed and discussed in detail in [4,5]. Some of them solve $F(x) = 0$ directly, whereas others solve the complementary piecewise linear equation System $H(z) = 0$. Note that there is a one-to-one solution correspondence between both representations [5]. Now, let us give an overview of some of these methods.

4.1 Full-Step Newton Variants

All continuous piecewise linear functions are known to be semi smooth. Hence the result in [10] ensures local convergence of the full-step iteration

$$x_+ = x - J^{-1}F(x), \text{ for } J \in \partial F(x)$$

to a solution x^*, provided that all generalized Jacobians $J \in \partial F(x^*)$ are nonsingular. However this condition need not be satisfied even if F is coherently oriented. Coherent orientation in some vicinity of x^* means that all limiting Jacobians $J \in \partial^L F(x^*)$ are nonsingular, so that the stronger result from [9], where the J are restricted to be limiting Jacobians, is applicable.

It should be noted that both results apply here in a trivial fashion, since convergence in one step must occur from all points x_0 belonging to polyhedra P_σ, whose closure contains x^*. Of course finding such an initial point x_0 requires to resolve all combinatorial issues in advance.

Hence we are more interested in global convergence results. We can guarantee full step convergence for the restricted generalized Newton method in finitely many steps towards the unique solution, if either of the contractivity conditions

$$\|I - J_\sigma^{-1} J_{\tilde{\sigma}}\| < 1, \text{ for all } \sigma, \tilde{\sigma} \text{ open}$$
$$\text{or} \qquad \|I - J_\sigma J_{\tilde{\sigma}}^{-1}\| < 1, \text{ for all } \sigma, \tilde{\sigma} \text{ open}$$

is satisfied w.r.t. to some induced matrix norm. The proof can be found in [5]. Either condition is rather strong and implies bijectivity. In terms of the abs-normal form they are implied by the conditions

$$\hat{\rho} \equiv \|Z\|\|J^{-1}Y\| < 1 - \|L\| \quad \text{and} \quad \frac{\hat{\rho}}{(1 - \hat{\rho} - \|L\|)(1 - \|L\|)} < \frac{1}{2}$$

As we have already noted suitable J_σ can be computed from the abs-normal form at reasonable expense.

Naturally the generalized Newton method with or without restriction to limiting Jacobians can also be applied to the complementary piecewise linear system, yielding

$$z_+ = z - (I - S\Sigma_z)^{-1} H(z) = (I - S\Sigma_z)^{-1} \hat{c}$$

However, here the local convergence condition that all limiting Jacobians be nonsingular is no weaker than the requirement that all generalized Jacobians be nonsingular. Sufficient for global full-step convergence are either of the following independent conditions

$$\|S\|_p < \tfrac{1}{3} \qquad \text{or} \qquad \rho(|S|) < \tfrac{1}{2}$$

where ρ denotes the spectral radius and $|S|$ the componentwise modulus.

If the second condition is satisfied, the calculation can be organized such that the whole solution process requires only $\frac{1}{3}s^3$ operations, just like a Gaussian elimination in the smooth linear case.

4.2 Piecewise Newton

Rather than taking full steps based on a local linearization one may restrict steps to stay within the closure of one polyhedron P_σ. This requires some pivoting and active set managament familiar from Lemke type algorithms for LCPs. For a comparitive study of the two approaches see the dissertation of T. Munson [7]. In [4] it was observed that coherent orientation implies, that the fibres

$$[x_0] \equiv \{x \in \mathbb{R}^n : F(x) = \lambda F(x_0),\ 0 < \lambda \in \mathbb{R}\}$$

are bifurcation-free piecewise linear paths for almost all $x_0 \in \mathbb{R}^n$. Then their closure contains a solution. Even in the case of singular fibres, there are strategies to reduce the residual towards a solution. An implementation is currently under development.

4.3 Modulus Algorithm

Checking F for surjectivity or openess is NP-hard, because there may be 2^n possible determinants $\det(J_\sigma)$, for $\sigma = \sigma_x$. An easier verifiable property is smooth dominance.

Definition 2. *$F : \mathbb{R}^n \to \mathbb{R}^n$ in abs-normal form is called smooth dominant, if for some nonsingular diagonal matrix D and a $p \in [1, \infty]$*

$$\|DSD^{-1}\|_p < 1$$

Smooth dominant abs-normal forms are always injective [5]. Nevertheless there are many practical problems which satisfy this condition.

In [2] Brugnano and Casulli consider unilateral constraints

$$\text{solve } \max(0, x) + Tx = -e/2$$

where $T \in \mathbb{R}^{n \times n}$ is an irreducible, symmetric, positive semidefinite matrix and $x, e \in \mathbb{R}^n$ vectors. This class of problems is piecewise linear and its abs-normal forms are smooth dominant. Electrical engineers considered piecewise linear function as models of electrical circuits since the 50's of the last century. For example Bokhoven discussed those models in his dissertation [1] and introduced the iteration

$$z_+ = S|z| - \hat{c}$$

whose convergence follows from smooth dominance, by the Banach fix point theorem. In our experience the modulus iteration is robust, but rather slow.

4.4 Alternating Block Seidel Iteration

Another fixed point iteration which has the potential of being significantly faster, is the following block Seidel scheme from [5]. Solving alternatingly the upper half for z and the lower half for x, we obtain $z_+ = h_z(h_x(z))$, where

$$\begin{aligned} h_z &: \mathbb{R}^n \to \mathbb{R}^s & h_z(x) &= (I - L\Sigma_x)^{-1}(c + Zx) \\ h_x &: \mathbb{R}^s \to \mathbb{R}^n & h_x(z) &= -J^{-1}b - J^{-1}Y\Sigma_z z \end{aligned}$$

The convergence of this method to the unique solution is ensured [5], if

$$\|S\|_p \leq \|L\|_p + \|ZJ^{-1}Y\|_p < 1$$

for some suitable p where positive diagonal scaling may be applied.

5 Conclusion and Outlook

We gave a short introduction to basic techniques of automatic differentiation and methods for the modelling of piecewise smooth functions via piecewise linearization with a second order error. We also discussed the solvabillity of the resulting equation systems in abs-normal form, by finitely convergent Newton variants or linearly convergent fix point solvers. Currently we are working on hybrid algorithms to obtain stable global and fast local convergence. They will then be used in the inner loop of a piecewise smooth equation solver by successive piecewise linearization. A related task to equation solving are the (un)constrained optimization of piecewise smooth objectives and the numerical integration of initial value problems with Lipschitzian right hand sides. Common utillities for manipulating abs-normal forms are developed as the linear algebra package PLAN-C, which uses abs-normal forms as objects.

References

1. van Bokhoven, W.M.G.: Piecewise-linear Modelling and Analysis. Kluwer Technische Boeken, The Netherlands (1981)
2. Brugnano, L., Casulli, V.: Iterative solution of piecewise linear systems. SIAM J. Sci. Comput. **30**(1), 463–472 (2008). SIAM
3. Griewank, A., Walther, A.: Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation. SIAM, Philadelphia (2008)
4. Griewank, A.: On stable piecewise linearization and generalized algorithmic differentiation. Optim. Methods Softw. **28**(6), 1139–1178 (2013)
5. Griewank, A., Bernt, J.-U., Radons, M., Streubel, T.: Solving piecewise linear equations in abs-normal form. optimization-online (2013)
6. Khan, K.A., Barton, P.I.: Evaluating an element of the Clarke generalized Jacobian of a piecewise differentiable function. In: Forth, S., et al. (eds.) Recent Advances in Algorithmic Differentiation, pp. 115–125. Springer, Heidelberg (2012)
7. Munson, T.S.: Algorithms and environments for complementarity. University of wisconsin, Diss. (2000)
8. Naumann, U.: The Art of Differentiating Computer Programs: An Introduction to Algorithmic Differentiation. SIAM, Philadelphia (2011)
9. Qi, L., Sun, D.: Nonsmooth equations and smoothing Newton methods. Applied Mathematics Report AMR 98.10 (1998)
10. Qi, L., Sun, J.: A nonsmooth version of Newton's method. Math. Prog. **58**(1–3), 353–367 (1993)
11. Rump, S.M.: Theorems of Perron-Frobenius type for matrices without sign restrictions. Linear Algebra Appl. **266**, 1–42 (1997)
12. Scholtes, S.: Introduction to Piecewise Differentiable Equations. Springer, New York (2012)

Efficient Smoothers for All-at-once Multigrid Methods for Poisson and Stokes Control Problems

Stefan Takacs(✉)

Mathematical Institute, University of Oxford, Oxford, UK
stefan.takacs@numa.uni-linz.ac.at
http://www.numa.uni-linz.ac.at/stefant/J3362/

Abstract. In the present paper we concentrate on an important issue in constructing a good multigrid solver: the choice of an efficient smoother. We will introduce all-at-once multigrid solvers for optimal control problems which show robust convergence in the grid size and in the regularization parameter. We will refer to recent publications that guarantee such a convergence behavior. These publications do not pay much attention to the construction of the smoother and suggest to use a normal equation smoother. We will see that using a Gauss Seidel like variant of this smoother, the overall multigrid solver is speeded up by a factor of about two with no additional work. The author will give a proof which indicates that also the Gauss Seidel like variant of the smoother is covered by the convergence theory. Numerical experiments suggest that the proposed method are competitive with Vanka type methods.

Keywords: PDE-constrained optimization · All-at-once multigrid · Gauss Seidel

1 Introduction

In the present paper we discuss the construction of the all-at-once multigrid solvers for two model problems. The first model problem is a standard *Poisson control problem*: Find a state $y \in H^1(\Omega)$ and a control $u \in L^2(\Omega)$ such that they minimize the cost functional

$$J(y,u) := \tfrac{1}{2}\|y - y_D\|^2_{L^2(\Omega)} + \tfrac{\alpha}{2}\|u\|^2_{L^2(\Omega)},$$

subject to the elliptic boundary value problem (BVP)

$$-\Delta y + y = u \text{ in } \Omega \qquad \text{and} \qquad \tfrac{\partial y}{\partial n} = 0 \text{ on } \partial\Omega.$$

The desired state y_D and the regularization parameter $\alpha > 0$ are assumed to be given. Here and in what follows, $\Omega \subseteq \mathbb{R}^2$ is a polygonal domain. We want to

The research was funded by the Austrian Science Fund (FWF): J3362-N25.

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 337–347, 2014.
DOI: 10.1007/978-3-662-45504-3_33

solve the finite element discretization of this problem using a fast linear solver which shows robust convergence behavior in the grid size and the regularization parameter. For solving this problem, we use the method of Lagrange multipliers, cf. [5,6]. We obtain a linear system in the state y, the control u and the Lagrange multiplier λ. In this linear system we eliminate the control as this has been done in [6,9]. We discretize the resulting system using the Courant element and obtain a linear system:

$$\underbrace{\begin{pmatrix} M_k & K_k \\ K_k & -\alpha^{-1} M_k \end{pmatrix}}_{\mathcal{A}_k :=} \underbrace{\begin{pmatrix} \underline{y}_k \\ \underline{\lambda}_k \end{pmatrix}}_{\underline{x}_k :=} = \underbrace{\begin{pmatrix} \underline{f}_k \\ 0 \end{pmatrix}}_{\underline{f}_k :=}. \tag{1}$$

Here, M_k and K_k are the standard mass and stiffness matrices, respectively. The control can be recovered using the following simple relation from the Lagrange multiplier: $\underline{u}_k = \alpha^{-1} \underline{\lambda}_k$, cf. [6]. In [6,12] it was shown that there are constants $\underline{C} > 0$ and $\overline{C} > 0$ (independent of the grid size h_k and the choice of α) such that the stability estimate

$$\| \mathcal{Q}_k^{-1/2} \mathcal{A}_k \mathcal{Q}_k^{-1/2} \| \leq \overline{C} \qquad \text{and} \qquad \| \mathcal{Q}_k^{1/2} \mathcal{A}_k^{-1} \mathcal{Q}_k^{1/2} \| \leq \underline{C}^{-1} \tag{2}$$

holds for the symmetric and positive definite matrix

$$\mathcal{Q}_k := \begin{pmatrix} M_k + \alpha^{1/2} K_k & \\ & \alpha^{-1} M_k + \alpha^{-1/2} K_k \end{pmatrix}.$$

The second model problem is a standard *Stokes control problem* (velocity tracking problem): Find a velocity filed $v \in [H^1(\Omega)]^d$, a pressure distribution $p \in L^2(\Omega)$ and a control $u \in [L^2(\Omega)]^d$ such that

$$J(v,p,u) = \tfrac{1}{2} \| v - v_D \|^2_{L^2(\Omega)} + \tfrac{\alpha}{2} \| u \|^2_{L^2(\Omega)}$$

is minimized subject to the Stokes equations

$$-\Delta v + \nabla p = u \text{ in } \Omega, \qquad \nabla \cdot v = 0 \text{ in } \Omega, \qquad v = 0 \text{ on } \partial\Omega.$$

The regularization parameter $\alpha > 0$ and the desired state (desired velocity field) $v_D \in [L^2(\Omega)]^d$ are assumed to be given. To enforce uniqueness of the solution, we additionally require $\int_\Omega p \, \mathrm{d}x = 0$.

Similar as above, we can set up the optimality system and eliminate the control, cf. [7,12]. The discretization can be done using the Taylor-Hood element. After these steps, we end up with the following linear system:

$$\underbrace{\begin{pmatrix} M_k & & K_k & D_k^T \\ & 0 & D_k & \\ K_k & D_k^T & -\alpha^{-1} M_k & \\ D_k & & & 0 \end{pmatrix}}_{\mathcal{A}_k :=} \underbrace{\begin{pmatrix} \underline{v}_k \\ \underline{p}_k \\ \underline{\lambda}_k \\ \underline{\mu}_k \end{pmatrix}}_{\underline{x}_k :=} = \underbrace{\begin{pmatrix} \underline{f}_k \\ 0 \\ 0 \\ 0 \end{pmatrix}}_{\underline{f}_k :=}. \tag{3}$$

where M_k and K_k are standard mass and stiffness matrices and D_k^T is the discretization of the gradient operator, see, e.g., [7,12]. Again, we are interested in a fast solver which is robust in the regularization parameter and the grid size. As in the previous example, the control $\underline{u}_k$ can by recovered from the Lagrange multiplier: $\underline{u}_k = \alpha^{-1}\underline{\lambda}_k$. In [12] it was shown that stability estimate (2) is satisfied for

$$\mathcal{Q}_k = \text{block-diag}\left(W_k,\ \alpha D_k W_k^{-1} D_k^T,\ \alpha^{-1} W_k,\ D_k W_k^{-1} D_k^T\right),$$

where $W_k := M_k + \alpha^{1/2} K_k$.

2 An All-at-once Multigrid Method

The linear systems (1) and (3) shall be solved by a multigrid method, which reads as follows. Starting from an initial approximation $\underline{x}_k^{(0)}$, one iterate of the multigrid method is given by the following two steps:

– *Smoothing procedure:* Compute

$$\underline{x}_k^{(0,m)} := \underline{x}_k^{(0,m-1)} + \hat{\mathcal{A}}_k^{-1}\left(\underline{f}_k - \mathcal{A}_k\, \underline{x}_k^{(0,m-1)}\right) \qquad \text{for } m = 1, \ldots, \nu$$

with $\underline{x}_k^{(0,0)} = \underline{x}_k^{(0)}$. The choice of the smoother (or, in other words, of the matrix $\hat{\mathcal{A}}_k^{-1}$) will be discussed below.

– *Coarse-grid correction:*

- Compute the defect $\underline{f}_k - \mathcal{A}_k\, \underline{x}_k^{(0,\nu)}$ and restrict it to grid level $k-1$ using an restriction matrix I_k^{k-1}: $\underline{r}_{k-1}^{(1)} := I_k^{k-1}\left(\underline{f}_k - \mathcal{A}_k\, \underline{x}_k^{(0,\nu)}\right)$.
- Solve the following coarse-grid problem approximatively:

$$\mathcal{A}_{k-1}\, \underline{p}_{k-1}^{(1)} = \underline{r}_{k-1}^{(1)} \tag{4}$$

- Prolongate $\underline{p}_{k-1}^{(1)}$ to the grid level k using an prolongation matrix I_{k-1}^k and add the result to the previous iterate: $\underline{x}_k^{(1)} := \underline{x}_k^{(0,\nu)} + I_{k-1}^k\, \underline{p}_{k-1}^{(1)}$.

As we have assumed to have nested spaces, the intergrid-transfer matrices can be chosen in a canonical way: I_{k-1}^k is the canonical embedding and the restriction I_k^{k-1} is its (properly scaled) transpose. If the problem (4) is solved exactly, we obtain the two-grid method. In practice, the problem (4) is approximatively solved by applying one step (V-cycle) or two steps (W-cycle) of the multigrid method, recursively. Only the coarsest grid level, (4) is solved exactly.

The only part of the multigrid algorithm that has not been specified yet, is the smoother. For the choice of the smoother, we make use of the convergence theory. We develop a convergence theory based on Hackbusch's splitting of the analysis into smoothing property and approximation property:

– *Smoothing property:*

$$\sup_{\underline{\tilde{x}}_k \in X_k} \frac{\left(\mathcal{A}_k(\underline{x}_k^{(0,\nu)} - \underline{x}_k^*), \underline{\tilde{x}}_k\right)_{\ell^2}}{\|\underline{\tilde{x}}_k\|_{\mathcal{L}_k}} \leq \eta(\nu)\|\underline{x}_k^{(0)} - \underline{x}_k^*\|_{\mathcal{L}_k} \tag{5}$$

should hold for some function $\eta(\nu)$ with $\lim_{\nu\to\infty} \eta(\nu) = 0$. Here and in what follows, $\underline{x}_k^* := \mathcal{A}_k^{-1}\underline{f}_k$ is the exact solution, $\|\cdot\|_{\mathcal{L}_k} := (\cdot,\cdot)_{\mathcal{L}_k}^{1/2} := (\mathcal{L}_k\cdot,\cdot)_{\ell^2}^{1/2}$ for some symmetric positive definite matrix $\mathcal{L}_k$ and $(\cdot,\cdot)_{\ell^2}$ is the standard Euclidean scalar product.

– *Approximation property:*

$$\|\underline{x}_k^{(1)} - \underline{x}_k^*\|_{\mathcal{L}_k} \leq C_A \sup_{\underline{\tilde{x}}_k \in X_k} \frac{\left(\mathcal{A}_k(\underline{x}_k^{(0,\nu)} - \underline{x}_k^*), \underline{\tilde{x}}_k\right)_{\ell^2}}{\|\underline{\tilde{x}}_k\|_{\mathcal{L}_k}}$$

should hold for some constant $C_A > 0$.

It is easy to see that, if we combine both conditions, we see that the two-grid method converges in the norm $\|\cdot\|_{\mathcal{L}_k}$ for ν large enough. The convergence of the W-cycle multigrid method can be shown under mild assumptions, see e.g. [3].

For the smoothing analysis, it is convenient to rewrite the smoothing property in pure matrix notation: (5) is equivalent to

$$\|\mathcal{L}_k^{-1/2}\mathcal{A}_k(I - \hat{\mathcal{A}}_k^{-1}\mathcal{A}_k)^\nu\mathcal{L}_k^{-1/2}\| \leq \eta(\nu). \tag{6}$$

For the Poisson control problem, it was shown in [6], that the approximation property is satisfied for the following choice of the matrix $\mathcal{L}_k$ (note that this matrix represents the norm $\|\cdot\|_{X^-}$ used in the mentioned paper)

$$\mathcal{L}_k = \begin{pmatrix} \operatorname{diag}(M_k + \alpha^{1/2}K_k) & \\ & \operatorname{diag}(\alpha^{-1}M_k + \alpha^{-1/2}K_k) \end{pmatrix},$$

i.e., $\mathcal{L}_k = \operatorname{diag}(\mathcal{Q}_k)$. Here and in what follows, $\operatorname{diag}(M)$ is the diagonal matrix containing the diagonal of a matrix M. For the Stokes control problem it was shown in [7], that the approximation property is satisfied for the following choice of $\mathcal{L}_k$:

$$\mathcal{L}_k = \begin{pmatrix} \hat{W}_k & & & \\ & \hat{P}_k & & \\ & & \alpha^{-1}\hat{W}_k & \\ & & & \alpha^{-1}\hat{P}_k \end{pmatrix},$$

where $\hat{W}_k := \operatorname{diag}(M_k + \alpha^{1/2}K_k)$ and $\hat{P}_k := \alpha\operatorname{diag}(D_k\hat{W}_k^{-1}D_k^T)$.

Still, we have not specified the choice of the smoother, which now can be done using the convergence theory. We have seen for which choices of $\mathcal{L}_k$ the approximation property is satisfied. We are interested in a smoother such that the smoothing property is satisfied for the same choice of $\mathcal{L}_k$.

In [7,9] a *normal equation smoother* was proposed. This approach is applicable to a quite general class of problems, cf. [2] and others. In our notation, the normal equation smoother reads as follows:

$$\underline{x}_k^{(0,m)} := \underline{x}_k^{(0,m-1)} + \tau \underbrace{\mathcal{L}_k^{-1} \mathcal{A}_k^T \mathcal{L}_k^{-1}}_{\hat{\mathcal{A}}_k^{-1} :=} \left(\underline{f}_k - \mathcal{A}_k \, \underline{x}_k^{(0,m-1)} \right) \quad \text{for } m = 1, \ldots, \nu.$$

Here, a fixed $\tau > 0$ has to be chosen such that the spectral radius $\rho(\tau \hat{\mathcal{A}}_k^{-1} \mathcal{A}_k)$ is bounded away from 2 on all grid levels k and for all choices of the parameters. It was shown that it is possible to find such an uniform τ for the Poisson control problem, e.g., in [9] and for the Stokes control problem, e.g., in [7]. For the normal equation smoother, the smoothing property can be shown using a simple eigenvalue analysis, cf. [2]. Numerical experiments show that the normal equation smoother works rather well for the mentioned model problems. However, there are smoothers such that the overall multigrid method converges much faster. Note that the normal equation smoother is basically a Richardson iteration scheme, applied to the normal equation. It is well-known for elliptic problems that Gauss Seidel iteration schemes are typically much better smoothers than Richardson iteration schemes. In the context of saddle point problems, the idea of Gauss Seidel smoothers has been applied, e.g., in the context of collective smoothers, see below. However, in the context of normal equation smoothers the idea of Gauss Seidel smoothers has not gained much attention. The setup of such an approach is straight forward: In compact notation such an approach, which we call *least squares Gauss Seidel* (LSGS) approach, reads as follows:

$$\underline{x}_k^{(0,m)} := \underline{x}_k^{(0,m-1)} + \underbrace{\mathrm{trig}(\mathcal{N}_k)^{-1} \mathcal{A}_k^T \mathcal{L}_k^{-1}}_{\hat{\mathcal{A}}_k :=} \left(\underline{f}_k - \mathcal{A}_k \, \underline{x}_k^{(0,m-1)} \right) \quad \text{for } m = 1, \ldots, \nu,$$

where $\mathcal{N}_k := \mathcal{A}_k^T \mathcal{L}_k^{-1} \mathcal{A}_k$ and $\mathrm{trig}(M)$ is a matrix whose coefficients coincide with the coefficients of M on the diagonal and the left-lower triangular part and vanish elsewhere. The author provides a possible realization of that approach as Algorithm 2 to convince the reader that the computational complexity of the LSGS approach is equal to the computational complexity of the normal equation smoother, where a possible realization is given as Algorithm 1.

We will see below that the LSGS approach works very well in the numerical experiments. However, there is no proof of the smoothing property known to the author. This is due to the fact that the matrix $\hat{\mathcal{A}}_k$ is not symmetric. One possibility to overcome this difficulty is to consider the symmetric version (symmetric least squares Gauss Seidel approach, sLSGS approach). This is analogous to the case of elliptic problems: For elliptic problems the smoothing property for the symmetric Gauss Seidel iteration can be shown for general cases but for the standard Gauss Seidel iteration the analysis is restricted to special cases, cf. Section 6.2.4 in [3].

One step of the sLSGS iteration consists of one step of the LSGS iteration, followed by one step of the LSGS iteration with reversed order of the variables.

Given: Iterate $(\mathtt{x}_i)_{i=1}^N = \underline{x}^{(0,m-1)}$ and corresp. residual $(\mathtt{r}_i)_{i=1}^N = \underline{f} - \mathcal{A}\underline{x}^{(0,m-1)}$;
Result: Iterate $(\mathtt{x}_i)_{i=1}^N = \underline{x}^{(0,m)}$ and corresp. residual $(\mathtt{r}_i)_{i=1}^N = \underline{f} - \mathcal{A}\underline{x}^{(0,m)}$;
for $i = 1, \dots, N$ **do**
  $\mathtt{q} := 0$;
  for all j **such that** $\mathcal{A}_{i,j} \neq 0$ **do** $\mathtt{q} := \mathtt{q} + \mathcal{A}_{i,j}/\mathcal{L}_{j,j} * \mathtt{r}_j$;
  $\mathtt{p}_i := \tau * \mathtt{q}/\mathcal{L}_{i,i}$;
end
for $i = 1, \dots, N$ **do**
  $\mathtt{x}_i := \mathtt{x}_i + \mathtt{p}_i$;
  for all j **such that** $\mathcal{A}_{j,i} \neq 0$ **do** $\mathtt{r}_j := \mathtt{r}_j - \mathcal{A}_{j,i} * \mathtt{p}_i$;
end

Algorithm 1. Normal equation iteration scheme

Given: Iterate $(\mathtt{x}_i)_{i=1}^N = \underline{x}^{(0,m-1)}$ and corresp. residual $(\mathtt{r}_i)_{i=1}^N = \underline{f} - \mathcal{A}\underline{x}^{(0,m-1)}$;
Result: Iterate $(\mathtt{x}_i)_{i=1}^N = \underline{x}^{(0,m)}$ and corresp. residual $(\mathtt{r}_i)_{i=1}^N = \underline{f} - \mathcal{A}\underline{x}^{(0,m)}$;
Prepare once: $\mathcal{N}_{i,i} := \sum_{j=1}^N \mathcal{A}_{i,j}^2/\mathcal{L}_{j,j}$ for all $i = 1, \dots, N$;
for $i = 1, \dots, N$ **do**
  $\mathtt{q} := 0$;
  for all j **such that** $\mathcal{A}_{i,j} \neq 0$ **do** $\mathtt{q} := \mathtt{q} + \mathcal{A}_{i,j}/\mathcal{L}_{j,j} * \mathtt{r}_j$;
  $\mathtt{p} := \mathtt{q}/\mathcal{N}_{i,i}$;
  $\mathtt{x}_i := \mathtt{x}_i + \mathtt{p}$;
  for all j **such that** $\mathcal{A}_{j,i} \neq 0$ **do** $\mathtt{r}_j := \mathtt{r}_j - \mathcal{A}_{j,i} * \mathtt{p}$;
end

Algorithm 2. LSGS iteration scheme

(So the computational complexity of one step of the sLSGS iteration is equal to the computational complexity of two steps of the standard LSGS iteration.) One step of the sLSGS iteration reads as follows in compact notation:

$$\underline{x}_k^{(0,m)} := \underline{x}_k^{(0,m-1)} + \hat{\mathcal{N}}_k^{-1} \mathcal{A}_k^T \mathcal{L}_k^{-1} \left(\underline{f}_k - \mathcal{A}_k \, \underline{x}_k^{(0,m-1)} \right) \qquad \text{for } m = 1, \dots, \nu,$$
$$\text{where } \hat{\mathcal{N}}_k := \mathrm{trig}(\mathcal{N}_k) \, \mathrm{diag}(\mathcal{N}_k)^{-1} \, \mathrm{trig}(\mathcal{N}_k)^T. \tag{7}$$

For our needs, the following convergence lemma is sufficient.

Lemma 1. *Assume that $\mathcal{A}_k$ is sparse, (2) is satisfied and let $\mathcal{L}_k$ be a positive definite diagonal matrix such that*

$$\|\mathcal{Q}_k^{1/2} \underline{x}_k\| \leq \|\mathcal{L}_k^{1/2} \underline{x}_k\| \quad \textit{for all } \underline{x}_k. \tag{8}$$

Then the sLSGS approach satisfies the smoothing property (6), i.e.,

$$\|\mathcal{L}_k^{-1/2} \mathcal{A}_k (I - \hat{\mathcal{N}}_k^{-1} \mathcal{N}_k)^\nu \mathcal{L}_k^{-1/2}\| \leq \frac{2^{-1/2} \, \overline{C} \; nnz(\mathcal{A}_k)^{5/2}}{\sqrt{\nu}},$$

where $nnz(M)$ is the maximum number of non-zero entries per row of M.

Note that (8) is a standard inverse inequality, which is satisfied for both model problems, cf. [6,7,9]. Note moreover that this assumption also has to be satisfied to show the smoothing property for the normal equation smoother, cf. [7,9].

Proof of Lemma 1. The combination of (2) and (8) yields $\|\mathcal{L}_k^{-1/2}\mathcal{A}_k\mathcal{L}_k^{-1/2}\| \leq \overline{C}$. Property 6.2.27 in [3] states that for any symmetric positive definite matrix $\mathcal{N}_k$

$$\|\hat{\mathcal{N}}_k^{-1/2}\mathcal{N}_k(I-\hat{\mathcal{N}}_k^{-1}\mathcal{N}_k)^\nu\hat{\mathcal{N}}_k^{-1/2}\| \leq \nu^{-1} \tag{9}$$

holds, where $\hat{\mathcal{N}}_k$ is as in (7). Using $\mathcal{D}_k := \mathrm{diag}(\mathcal{N}_k)$, we obtain

$$\begin{aligned}\|\mathcal{L}_k^{-1/2}\hat{\mathcal{N}}_k^{1/2}\|^2 &= \rho(\mathcal{L}_k^{-1/2}\hat{\mathcal{N}}_k\mathcal{L}_k^{-1/2}) \leq \|\mathcal{L}_k^{-1/2}\mathrm{trig}(\mathcal{N}_k)\mathcal{D}_k^{-1/2}\|^2\\ &\leq \|\mathcal{L}_k^{-1/2}\mathcal{D}_k^{1/2}\|^2\|\mathcal{D}_k^{-1/2}\mathrm{trig}(\mathcal{N}_k)\mathcal{L}_k^{-1/2}\|^2\end{aligned}$$

Let $\mathcal{A}_k = (\mathcal{A}_{i,j})_{i,j=1}^N$, $\mathcal{N}_k = (\mathcal{N}_{i,j})_{i,j=1}^N$, $\mathcal{L}_k = (\mathcal{L}_{i,j})_{i,j=1}^N$ and $\psi(i) := \{j \in \mathbb{N} : \mathcal{N}_{i,j} \neq 0\}$. We obtain using Gerschgorin's theorem, the fact that the infinity norm is monotone in the matrix entries, and using the symmetry of $\mathcal{N}_k$ and $\mathcal{A}_k$ and Cauchy-Schwarz inequality:

$$\begin{aligned}&\|\mathcal{D}_k^{-1/2}\mathrm{trig}(\mathcal{N}_k)\mathcal{D}_k^{-1/2}\|\\ &\leq \|\mathcal{D}_k^{-1/2}\mathrm{trig}(\mathcal{N}_k)\mathcal{D}_k^{-1/2}\|_\infty^{1/2}\|\mathcal{D}_k^{-1/2}\mathrm{trig}(\mathcal{N}_k)^T\mathcal{D}_k^{-1/2}\|_\infty^{1/2} \leq \|\mathcal{D}_k^{-1/2}\mathcal{N}_k\mathcal{D}_k^{-1/2}\|_\infty\\ &= \max_{i=1,\dots,N}\sum_{k\in\psi(i)}\left(\sum_{n=1}^N\frac{\mathcal{A}_{i,n}^2}{\mathcal{L}_{n,n}}\right)^{-1/2}\left(\sum_{j=1}^N\frac{\mathcal{A}_{i,j}\mathcal{A}_{j,k}}{\mathcal{L}_{j,j}}\right)\left(\sum_{n=1}^N\frac{\mathcal{A}_{k,n}^2}{\mathcal{L}_{n,n}}\right)^{-1/2}\\ &\leq \max_{i=1,\dots,N}\sum_{k\in\psi(i)}1 = \mathrm{nnz}(\mathcal{N}_k) \leq \mathrm{nnz}(\mathcal{A}_k)^2.\end{aligned} \tag{10}$$

Further, we obtain

$$\begin{aligned}\|\mathcal{L}_k^{-1/2}\mathcal{D}_k^{1/2}\|^2 = \|\mathcal{L}_k^{-1/2}\mathcal{D}_k^{1/2}\|_\infty^2 = \|\mathcal{L}_k^{-1/2}\mathcal{D}_k\mathcal{L}_k^{-1/2}\|_\infty = \max_{i=1,\dots,N}\sum_{j=1}^N\frac{\mathcal{A}_{i,j}^2}{\mathcal{L}_{i,i}\mathcal{L}_{j,j}}\\ \leq \mathrm{nnz}(\mathcal{A}_k)\max_{i,j=1,\dots,N}\frac{\mathcal{A}_{i,j}^2}{\mathcal{L}_{i,i}\mathcal{L}_{j,j}} = \mathrm{nnz}(\mathcal{A}_k)\|\mathcal{L}^{-1/2}\mathcal{A}\mathcal{L}^{-1/2}\|^2 \leq \mathrm{nnz}(\mathcal{A}_k)\,\overline{C}^2.\end{aligned} \tag{11}$$

By combining (9), (10) and (11), we obtain

$$\begin{aligned}&\|\mathcal{L}_k^{-1/2}\mathcal{A}_k(I-\hat{\mathcal{N}}_k^{-1}\mathcal{N}_k)^\nu\mathcal{L}_k^{-1/2}\|^2\\ &\quad\leq \|\mathcal{L}_k^{-1/2}(I-\mathcal{N}_k\hat{\mathcal{N}}_k^{-1})^\nu\mathcal{A}_k\mathcal{L}_k^{-1}\mathcal{A}_k(I-\hat{\mathcal{N}}_k^{-1}\mathcal{N}_k)^\nu\mathcal{L}_k^{-1/2}\|\\ &\quad= \|\mathcal{L}_k^{-1/2}\mathcal{N}_k(I-\hat{\mathcal{N}}_k^{-1}\mathcal{N}_k)^{2\nu}\mathcal{L}_k^{-1/2}\| \leq \frac{\overline{C}^2\mathrm{nnz}(\mathcal{A}_k)^5}{2\nu},\end{aligned}$$

which finishes the proof. □

We went to compare the numerical behavior of the LSGS approach with the behavior of a standard smoother. One class of standard smoothers for saddle point problems is the class of *Vanka type smoothers*, which has been originally introduced for Stokes problems, cf. [11]. Such smoothers have also gained interest for optimal control problems, see, e.g., [1,8,10].

The idea of Vanka type smoothers is to compute updates in subspaces directly for the whole saddle point problem and to combine these updates is an additive or a multiplicative way to compute the next update. Here, the variables are not grouped based on the block-structure of $\mathcal{A}_k$, but the grouping is done of based on the location of the corresponding degrees of freedom in the domain Ω. The easiest of such ideas for the Poisson control problems is to do the grouping point-wise, which leads to the idea of *point smoothing*. Here, we group for each node δ_i of the discretization (each degree of freedom of the Courant element) the value y_i of the state and the value λ_i of the Lagrange multiplier and compute an update in the corresponding subspace. The multiplicative variant of such a smoother is a *collective Gauss Seidel* (CGS) smoother:

$$\underline{x}_k^{(0,m,i)} := \underline{x}_k^{(0,m,i-1)} + \mathcal{P}_k^{(i)} \left(\mathcal{P}_k^{(i)\,T} \mathcal{A}_k \mathcal{P}_k^{(i)} \right)^{-1} \mathcal{P}_k^{(i)\,T} \left(\underline{f}_k - \mathcal{A}_k \, \underline{x}_k^{(0,m,i-1)} \right),$$

where $\underline{x}_k^{(0,m,0)} := \underline{x}_k^{(0,m-1)}$ and $\underline{x}_k^{(0,m)} := \underline{x}_k^{(0,m,N_k)}$. For each $i = 1, \ldots, N_k$, the matrix $\mathcal{P}_k^{(i)} \in \mathbb{R}^{2N_k \times 2}$ takes the value 1 on the positions $(i,1)$ and $(i+N_k,2)$ and the value 0 elsewhere. For the Poisson control problem, we obtain

$$\mathcal{P}_k^{(i)\,T} \mathcal{A}_k \mathcal{P}_k^{(i)} = \begin{pmatrix} M_{i,i} & K_{i,i} \\ K_{i,i} & -\alpha^{-1} M_{i,i} \end{pmatrix},$$

where $M_{i,i}$ and $K_{i,i}$ are the entries of the matrices M_k and K_k.

For the Stokes control problem, it is not reasonable to use exactly the same approach. This is basically due to the fact that the degrees of freedom for v and λ are not located on the same positions as the degrees of freedom for p and μ. However, we can introduce an approach based on patches: so, for each vertex of the triangulation, we consider subspaces that consist of the degrees of freedoms located on the vertex itself and the degrees of freedom located on all edges which have one end at the chosen vertex, cf. Fig. 1. Note that here the subspaces are

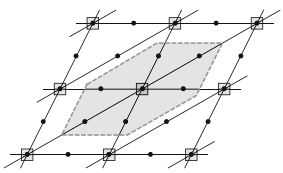


Fig. 1. Patches for the Vanka-type smoother applied to a Taylor Hood discretization. The dots are the degrees of freedom of v and λ, the rectangles are the degrees of freedom of p and μ

much larger than the subspaces chosen in the case of the CGS approach for the Poisson control problem (which was just 2). This increases the computational cost of applying the method significantly. For Vanka type smoothers there are only a few convergence results known, cf. [1] for a Fourier Analysis and an analysis based compactness argument and [8] for a proof based on Hackbusch's splitting of the analysis into smoothing property and approximation property which shows the convergence in case of a collective Richardson smoother.

3 Numerical Results

In this section we give numerical results to illustrate quantitatively the convergence behavior of the proposed methods. The number of iterations was measured as follows: We start with a random initial guess and iterate until the relative error in the norm $\|\cdot\|_{\mathcal{L}_k}$ was reduced by a factor of 10^{-6}. Without loss of generality, the right-hand side was chosen to be 0. For both model problems, the normal equation smoother, the LSGS smoother, the sLSGS smoother and a Vanka type smoother have been applied. For the smoothers 2 pre- and 2 post-smoothing steps have been applied. Only for the sLSGS smoother, just 1 pre- and 1 post-smoothing step has been applied. This is due to the fact that one step of the symmetric version is basically the same computational cost as two steps of the standard version. The normal equation smoother was damped with $\tau = 0.4$ for the Poisson control problem and $\tau = 0.35$ for the Stokes control problem, cf. [7,9]. For the Gauss Seidel-like approaches, damping was not used.

In Table 1, we give the results for the standard Poisson control problem. Here, we see that all smoothers lead to convergence rates that are well bounded for a wide range of h_k and α. Compared to the normal equation smoother, the LSGS smoother leads to a speedup be a factor of about two without any additional work. The symmetric version (sLSGS) is a bit slower than the LSGS method. For the first model problem, the (popular) CGS method is significantly faster. However, for this method no convergence theory is known.

Table 1. Number of iterations for the *Poisson control model problem*

	Normal equation			LSGS			sLSGS			CGS		
$\alpha =$	10^0	10^{-6}	10^{-12}	10^0	10^{-6}	10^{-12}	10^0	10^{-6}	10^{-12}	10^0	10^{-6}	10^{-12}
$k = 5$	26	31	28	11	9	7	14	12	14	5	5	3
$k = 6$	27	28	29	11	11	7	14	14	13	5	5	3
$k = 7$	27	28	31	11	11	6	14	14	12	5	5	3
$k = 8$	27	27	25	11	11	3	14	14	7	5	5	4

In Table 2, we give the convergence results for the Stokes control problem. Also here we observe that the LSGS and the sLSGS approach lead to a speedup of a factor of about two compared to the normal equation smoother. Here, the

Table 2. Number of iterations for the *Stokes control model problem*

	Normal equation			LSGS			sLSGS			Vanka type		
$\alpha =$	10^0	10^{-6}	10^{-12}	10^0	10^{-6}	10^{-12}	10^0	10^{-6}	10^{-12}	10^0	10^{-6}	10^{-12}
$k = 4$	31	31	60	13	12	14	17	16	22	11	10	7
$k = 5$	32	30	55	14	13	12	18	16	19	11	10	7
$k = 6$	32	31	44	14	13	9	18	17	12	11	11	7
$k = 7$	32	31	37	14	14	6	18	17	9	11	11	9

Vanka type smoother shows slightly smaller iteration numbers than the LSGS approach. In terms of computational costs, the LSGS smoother seems to be much better than the patch-based Vanka type smoother because there relatively large subproblems have to be solved to compute the updates. This is different the case of the CGS smoother, where the subproblems are just 2-by-2 linear systems. Numerical experiments have shown that the undamped version of the patch-based Vanka type method does not lead to a convergent multigrid method. So, this smoother was damped with $\tau = 0.4$. Due to lack of convergence theory, the author cannot explain why this approach – although it is a multiplicative approach – needs damping.

For completeness, the author wants to mention that for cases, where a (closed form of a) matrix $\mathcal{Q}_k$ satisfying (2) robustly is not known, the normal equation smoother does not show as good results as methods where such an information is not needed, like Vanka type methods. This was discussed in [8] for a boundary control problem, but it is also true for the linearization of optimal control problems with inequality constraints as discussed in [4] and others. The same is true for the Gauss Seidel like variants of the normal equation smoother.

Concluding, we have observed that accelerating the idea of normal equation smoothing with a Gauss Seidel approach, leads to a speedup of a factor of about two without any further work. The fact that convergence theory is known for the sLSGS approach, helps also for the numerical practice (unlike the case of Vanka type smoothers).

References

1. Borzi, A., Kunisch, K., Kwak, D.Y.: Accuracy and convergence properties of the finite difference multigrid solution of an optimal control optimality system. SIAM J. Control Optimization **41**(5), 1477–1497 (2003)
2. Brenner, S.C.: Multigrid methods for parameter dependent problems, RAIRO. Modélisation Math. Anal. Numér **30**, 265–297 (1996)
3. Hackbusch, W.: Multi-Grid Methods and Applications. Springer, Berlin (1985)
4. Herzog, R., Sachs, E.: Preconditioned conjugate gradient method for optimal control problems with control and state constraints. SIAM J. Matrix Anal. Appl. **31**(5), 2291–2317 (2010)
5. Lions, J.L.: Optimal Control of Systems Governed by Partial Differential Equations. Springer, Heidelberg (1971)

6. Schöberl, J., Simon, R., Zulehner, W.: A robust multigrid method for elliptic optimal control problems. SIAM J. Numerical Anal. **49**, 1482–1503 (2011)
7. Takacs, S.: A robust all-at-once multigrid method for the Stokes control problem (2013) (submitted)
8. Takacs, S., Zulehner, W.: Convergence analysis of multigrid methods with collective point smoothers for optimal control problems. Comput. Vis. Sci. **14**(3), 131–141 (2011)
9. Takacs, S., Zulehner, W.: Convergence analysis of all-at-once multigrid methods for elliptic control problems under partial elliptic regularity. SIAM J. Numerical Anal. **51**(3), 1853–1874 (2013)
10. Trottenberg, U., Oosterlee, C., Schüller, A.: Multigrid. Academic Press, London (2001)
11. Vanka, S.P.: Block-implicit multigrid solution of Navier-Stokes equations in primitive variables. Math. Comp. **65**, 138–158 (1986)
12. Zulehner, W.: Non-standard norms and robust estimates for saddle point problems. SIAM J. Matrix Anal. Appl. **32**, 536–560 (2011)

Continuous-Time Local Model Network for the Boost-Pressure Dynamics of a Turbocharger

Christoph Weise[1], Kai Wulff[1(✉)], Marc-Hinrik Höper[2(✉)], and Romain Hurtado[2]

[1] Control Engineering Group, TU Ilmenau, Ilmenau, Germany
kai.wulff@tu-ilmenau.de
[2] Engine Systems, IAV GmbH, Gifhorn, Germany
marc-hinrik.hoeper@iav.de

Abstract. In this paper we consider continuous-time local model networks (LMN) to model dynamical processes with strong nonlinearities. The local model approach allows for simple black-box identification procedures using experimental data. Using the LoLiMoT algorithm the number of models can be significantly reduced and may yield insights into the nonlinearities driving the process. We propose a variation of the LoLiMoT algorithm that partitions the operating range in a more efficient manner and proves particular suited for heterogenous nonlinearities.

Keywords: Local model network · Modelling · Turbocharger

1 Introduction

Mathematical modelling is an essential tool for system analysis and control. In this regard we motivate our aim to derive a mathematical model that describes the considered process in an adequate fashion. Precise physical models, however, are often hard to obtain and are frequently challenged by tedious parameter estimation. Moreover, physical models can be very sensitive with respect to parameter variations such that an appropriate set of parameters is very hard to obtain. In particular nonlinear dynamical processes suffer from this problem. Resorting to grey-box models that are driven by experimental data can often resolve this problem. For processes with nonlinear dynamics local-model networks can be used to represent the global dynamics by identifying simple linear models locally and combine them to a network that matches the nonlinear global dynamics sufficiently. This is particular useful in industrial practice where simple controller structures with few tunable parameters are often preferred. Therefore we aim for an continuous-time LMN such that most common manual control-design methods are readily applicable.

The basic concepts for local model networks have been developed in the nineties, see e.g. [1,2], for early work. Ever since local model networks have been

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. 348–358, 2014.
DOI: 10.1007/978-3-662-45504-3_34

studied for various system classes and applications range from mechatronical systems to process engineering. Typical issues that arise in most studies pertain to the implementation of the local model network. In [3,4] the local models are realised each having their own local state-space. The overall output of the model network is then obtained by interpolation of the local model outputs. The complement approach is taken in [5], where the state is shared by all local models. Here the parameters are interpolated in appropriate fashion to emulate the nonlinear dynamics of the process. The vast majority of the publications, however, consider discrete-time systems.

An important role for the complexity of the local model network plays the choice of the operating regimes of the individual models. In this regard several methods have been proposed to partition the scheduling space. A summary on existing partitioning strategies is given in [6,7].

More details on local model trees based on recursive orthogonal splitting (LoLiMoT) are given in [2,8–11]. This partitioning strategy uses hyper-rectangles and is time-consuming regarding large scheduling problems. Axis oblique partitioning strategies using the HiLoMoT algorithm can be found in [6,12]. Other publications focus on the optimisation of the partitions for a fixed number of models [5] itself. Based on the chosen partitioning the used weighting functions have an additional effect on the performance of the LMN [13].

In this contribution we investigate the suitability of local model networks for continuous-time domain. In regard to the controller design the dynamics and the gain of the process are equally important. In particular whenever the dominating time-constant of the dynamics vary within the operating range, good knowledge of the latter is vital in order to design high-performing controllers. Therefore we choose an approach that is able to map individual dynamics to different points in the scheduling space, as well as individual gains.

For the resulting local model network we address in particular issues regarding the offsets of the local models and suitable partitioning techniques. For processes that exhibit strong nonlinear dynamics, the local model network may need a very large number of models to match the behaviour appropriately, when the local operating points are evenly distributed. In order to keep the number of local models moderately large, we apply the LoLiMoT algorithm and propose a novel technique that is able to introduce multiple partitions in each iteration and may lead to partitions that fit the nonlinearity better.

The described techniques are applied to model the boots-pressure dynamics of a turbo charged combustion engine. The experimental data for modelling and verification are obtained from a full-scale test rig.

2 Local Model Networks

The local model network (LMN) approach [1] uses the strategy of divide and conquer to describe a complex and non-linear behaviour of a dynamical system, given by:

$$\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x}(t), \boldsymbol{u}(t), t)\,. \tag{1}$$

First of all the operating range of the system is divided in M regions, wherein the dynamics can be approximated by a more simple (e.g. linear or affine) model $\tilde{\boldsymbol{h}}_i(\boldsymbol{x}(t), \boldsymbol{u}(t), t)$.

In order to cover the complete operating range the interaction of the local models is controlled by the scheduler $\boldsymbol{\Phi}$, a vector of exogenous or system variables prior chosen. Each model is weighted a function by $\rho_i(\boldsymbol{\Phi}) \in [0, 1]$, representing the range of validity of the i-th model. The system dynamics (1) are then approximated by the LMN by the interpolation of the local models

$$\tilde{\boldsymbol{y}} = \sum_{i=1}^{M} \rho_i(\boldsymbol{\Phi})\ \tilde{\boldsymbol{h}}_i\big(\boldsymbol{x}(t), \boldsymbol{u}(t), t\big) \qquad \text{with} \qquad \sum_{i=1}^{M} \rho_i(\boldsymbol{\Phi}) = 1\,.$$

In this paper we examine the application of the LMN approach to nonlinear dynamical systems of the form

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t)), \quad \boldsymbol{x}(0) = \boldsymbol{x}_0 \quad \text{with:} \quad \boldsymbol{x} \in \mathbb{R}^m, \boldsymbol{y} \in \mathbb{R}^q, \boldsymbol{u} \in \mathbb{R}^p \\ \boldsymbol{y}(t) &= \boldsymbol{h}(\boldsymbol{x}(t), \boldsymbol{u}(t))\,. \end{aligned} \tag{2}$$

The nonlinear dynamics shall be approximated by linearisations at M stationary solutions $(\boldsymbol{x}_i^*, \boldsymbol{u}_i^*)$ with $\boldsymbol{f}(\boldsymbol{x}_i^*, \boldsymbol{u}_i^*) = 0$:

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{f}(\boldsymbol{x}_i^*, \boldsymbol{u}_i^*) + \boldsymbol{A}_i(\boldsymbol{x}(t) - \boldsymbol{x}_i^*) + \boldsymbol{B}_i(\boldsymbol{u}(t) - \boldsymbol{u}_i^*) \\ \boldsymbol{y}(t) &= \boldsymbol{h}(\boldsymbol{x}_i^*, \boldsymbol{u}_i^*) + \boldsymbol{C}_i(\boldsymbol{x}(t) - \boldsymbol{x}_i^*) + \boldsymbol{D}_i(\boldsymbol{u}(t) - \boldsymbol{u}_i^*)\,. \end{aligned} \tag{3}$$

Following [3] we can define M models with local states $\tilde{\boldsymbol{x}}_i(t) = \boldsymbol{x}(t) - \boldsymbol{x}_i^*$, inputs $\tilde{\boldsymbol{u}}_i(t) = \boldsymbol{u}(t) - \boldsymbol{u}_i^*$ and outputs $\tilde{\boldsymbol{y}}_i(t) = \boldsymbol{y}(t) - \boldsymbol{h}(\boldsymbol{x}_i^*, \boldsymbol{u}_i^*)$. This leads to a local-state representation of the LMN where the output is a weighted sum of the single models' outputs. This description proves sufficient for controller design based on local observes as discussed in [3].

However, simulation purposes or even stability analysis may pose further requirements onto the LMN [14,15]. In a local-state architecture, the initial state of the local model may induce strong transient responses at the switching instance which are not desirable and may affect stability [16]. Furthermore, whenever the scheduling variable switches different output signals between the subsequent local models will cause discontinuities of the output, which will not be observed in the original nonlinear process behaviour.

Therefore we shall choose a global-state representation of the LMN, where the state-vector $\boldsymbol{x}$ is shared by all local models. Then the linearisation (3) becomes an affine system with local system-offset $\boldsymbol{K}_i$ and the local output-offset $\boldsymbol{L}_i$:

$$\begin{aligned} \boldsymbol{K}_i &= -(\boldsymbol{A}_i\boldsymbol{x}_i^* + \boldsymbol{B}_i\boldsymbol{u}_i^*) \\ \boldsymbol{L}_i\ &= \boldsymbol{h}(\boldsymbol{x}_i^*, \boldsymbol{u}_i^*) - (\boldsymbol{C}_i\boldsymbol{x}_i^* + \boldsymbol{D}_i\boldsymbol{u}_i^*)\,. \end{aligned} \tag{4}$$

For the global-state LMN we obtain the parameter-varying affine system

$$\begin{aligned} \dot{\boldsymbol{x}} &= \boldsymbol{A}(\boldsymbol{\Phi})\boldsymbol{x} + \boldsymbol{B}(\boldsymbol{\Phi})\boldsymbol{u} + \boldsymbol{K}(\boldsymbol{\Phi}) \\ \boldsymbol{y} &= \boldsymbol{C}(\boldsymbol{\Phi})\boldsymbol{x} + \boldsymbol{D}(\boldsymbol{\Phi})\boldsymbol{u} + \boldsymbol{L}(\boldsymbol{\Phi}). \end{aligned} \tag{5}$$

where $\boldsymbol{A}(\boldsymbol{\Phi}) = \sum_{i=1}^{M} \rho_i(\boldsymbol{\Phi})\boldsymbol{A}_i$ and $\boldsymbol{B}(\boldsymbol{\Phi})$, $\boldsymbol{K}(\boldsymbol{\Phi})$, $\boldsymbol{C}(\boldsymbol{\Phi})$, $\boldsymbol{D}(\boldsymbol{\Phi})$ and $\boldsymbol{L}(\boldsymbol{\Phi})$ with according interpolations.

Note that the state-space of the global-state LMN has M-time smaller dimension than that of the local-state LMN. Furthermore, the state is continuous for all variations in $\boldsymbol{\Phi}$. With some mild assumptions the same holds for the output.

3 Experimental Identification of the Pressure-Dynamics

Modern combustion engines are frequently equipped with an exhaust turbocharger that boosts the pressure of the intake air in order to increase the amount of oxygen in the cylinder. Often the compressor power can be manipulated by varying the geometry of the turbine, so-called variable-geometry turbine (VGT). This allows for influencing the charging pressure within certain constraints. While the physical relations of the quantities are well known, e.g. [17], deriving a dynamical model of the pressure dynamics and identifying its parameters is very time-consuming as such models exhibit strong non-linearities and are very sensitive with respect to certain parameter variations [18–20]. A data-driven black-box approach may therefore prove valuable in this context and, thus, the pressure dynamics are a very suitable process to investigate properties and challenges of the LMN approach.

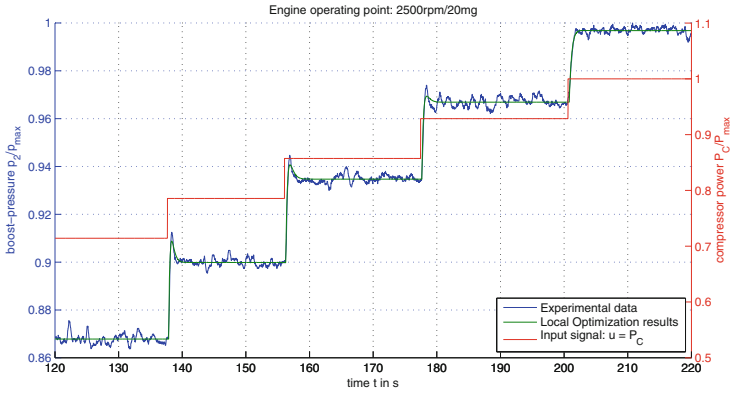


Fig. 1. Sample of data time-series and identification result

The data-base for identification consists of 16 time-series of experimental input-output data at various engine operating points (n_i, q_i). Within one test-cycle the engine speed n_i and injection rate q_i was maintained constant whereas the compressor power P_C was increased stepwise, see Fig. 1 for a sample time-series.

3.1 LMN Identification

The engine speed and injection rate are natural scheduling variables from an engineering point of view. We can further observe that the dynamics vary significantly for different levels of the compressor power P_C, cf. Fig. 1. Therefore we choose $\boldsymbol{\Phi} = [n, q, P_C]^T$ as scheduling vector; the system's input is the compressor power $u := P_C$, the output is the boost-pressure $y := p_2$.

Based on the data structure the most simple partitioning approach of a regular grid is used to divide the total working range into 256 local regions. For each of these local regions we choose a grey-box approach to identify the local dynamics. In our case physical considerations lead to the following parametrisation of the local models:

$$G(s) = V \frac{\tau_1 s + 1}{(\tau_2 s + 1)(\tau_3 s + 1)} .$$

The parameters $(V_i, \tau_{1i}, \tau_{2i}, \tau_{3i})$ of the local models $i = 1, \ldots, M$ are obtained by non-linear optimisation for each individual step in the time-series with removed offsets. In order to cast the local models in into the state-space representation (5) we choose the observer canonical form which has a constant output matrix that renders the output continuous for all variations of $\boldsymbol{\Phi}$. The regions of validity of each model are defined by triangular weighting functions ρ_i.

The offset parameters $\boldsymbol{K}_i$ and L_i in (4) are chosen to account for the removed offset in the local identification. Without loss of generality we choose $L_i = 0$ for all i. For calculating $\boldsymbol{K}_i$ we note, that the second state-component is the output and thus can be taken from the data. For the first local model of each time-series we have:

$$\boldsymbol{K}_i = -\boldsymbol{B}_i u(0) - \boldsymbol{A}_i \begin{pmatrix} x_1(0) \\ y(0) \end{pmatrix} ,$$

where only $x_1(0)$ can be chosen arbitrarily. For the second step of a given time-series we obtain the initial condition from the final state using the already identified parameters of the first model, etc. Determining the offsets in this fashion yields a continuous output for each time-series as shown in Fig. 1.

3.2 Verification of the LMN

In order to verify the LMN's ability to match the non-linear process dynamics, we simulate the LMN using a highly dynamic verification-cycle covering the full operation range. Figure 2 shows the evolution of the scheduling variable's components during that test-cycle featuring steep sloped jumps as well as various smooth transitions.

The overall performance of the LMN is quite satisfactory with a mean square error of $MSE = 2.72 \times 10^3\,\mathrm{hPa}^2$ and a maximum error $|e_{max}| = 223.6\,\mathrm{hPa}$. Figure 3 shows two details of the full cycle. The model matches the experimental data very well for dynamical (right) and stationary (left) areas. However, we note that for high compressor powers the stationary error may rise as high as 100 hPa. This may be due to high noise-levels in the data for these operating points that prohibit a precise estimate of the DC-gain in these regions.

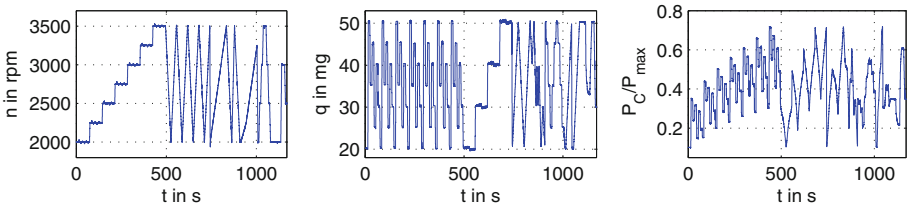


Fig. 2. Evolution of the scheduling variables during the validation cycle.

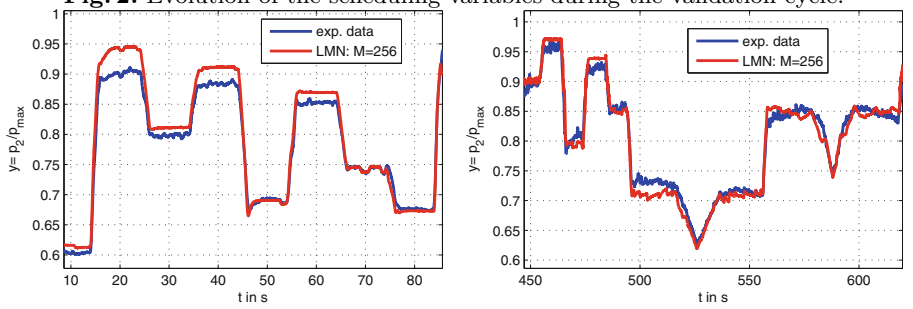


Fig. 3. Simulation result using the validation cycle with 256 local models.

4 Local Model Trees

The grid-based approach of the previous section typically leads to a large number of local models that are equally spaced in the operating range, irrespective of the character of the nonlinearity. Local model tree algorithms address these issues by starting with a low number of models and improving the model network iteratively by introducing additional models, if the error is large in a certain region. Thereby, only few models will be placed in regions where the process behaves in an almost linear fashion and local over-fitting is avoided.

While there are a number of partitioning strategies available, an orthogonal partitioning appears to be a natural choice in our case, as the data for identification is distributed in a grid-like fashion within the scheduling space. The classical LoLiMoT-algorithm presented in [9,10] uses such orthogonal partitioning. In each iteration the region of the worst performing local model is split in 2 hyper-rectangles wherein the parameters for new local models are identified. Every possible division (one for each dimension of the scheduling space) is analysed and the best division is chosen.

4.1 Local Error Based Partitioning (LEB)

The classical LoLiMoT considers the accumulated quadratic error to evaluate the performance of the model and thus does not use the full available information at each iteration. In this section we propose a novel partitioning strategy that uses the signed error at each point in the scheduling space in relation to certain thresholds τ^- and τ^+ dividing the scheduling space into 3 regions in one iteration. As illustrated in Fig. 4 this may lead to a better approximation of

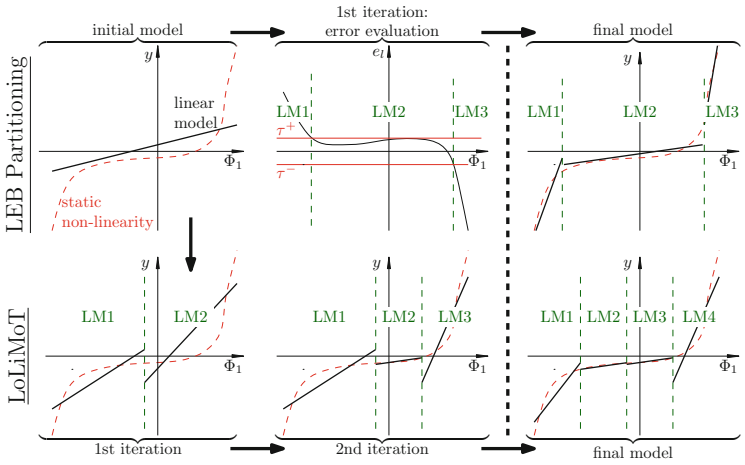


Fig. 4. Illustration of local error based partitioning (top) and classical LoLiMoT partitioning (bottom) for a simple static nonlinearity.

the nonlinearity by using a smaller number of models compared to the classical bisection.

Our optimisation data is given at discrete points of the scheduling space. Therefore we can define the local error

$$e_l(\boldsymbol{\Phi}_k) := \int_0^{t_{\text{end}}} [\hat{y}(t) - y(t)]\, \rho_k\left(\boldsymbol{\Phi}(t)\right) \mathrm{d}t$$

for each local model $k = 1, 2, \ldots, M_{max}$. In order to find the best cut in the scheduling space the local error is projected onto one component by the weighted sum of the local errors of the remaining dimensions.

For the boost-pressure dynamics we obtained either a quadratic or linear distribution of this projected local error (PLE), cf. Fig. 5. Based on this classification of the PLE distribution an adaptive threshold τ is calculated:

1. If the PLE shows a parabolic distribution (with 2 zero-crossings) the threshold is set to zero $\tau = 0$ defining three regions, based on the requirement that the area of validity of a local model should be compact.
2. If the PLE shows a linear distribution (with a single zero-crossing) the symmetric thresholds $|\tau^+| = |\tau^-|$ are calculated based on the PLE: $\tau^{\pm} = \pm\kappa \cdot \min\{e_{l,min}; e_{l,max}\}$ with a constant factor $\kappa < 1$ three regions are set.
3. If the PLE distribution cannot be categorised clearly, the classical bisection is applied defining two new regions.

Finally we embedded this approach into the LoLiMoT algorithm [9] (see Fig. 5):

1. The weighting functions of the initial model are calculated. If only one global model is used, the global weight is set to $\rho_1(\boldsymbol{\Phi}) = 1$.

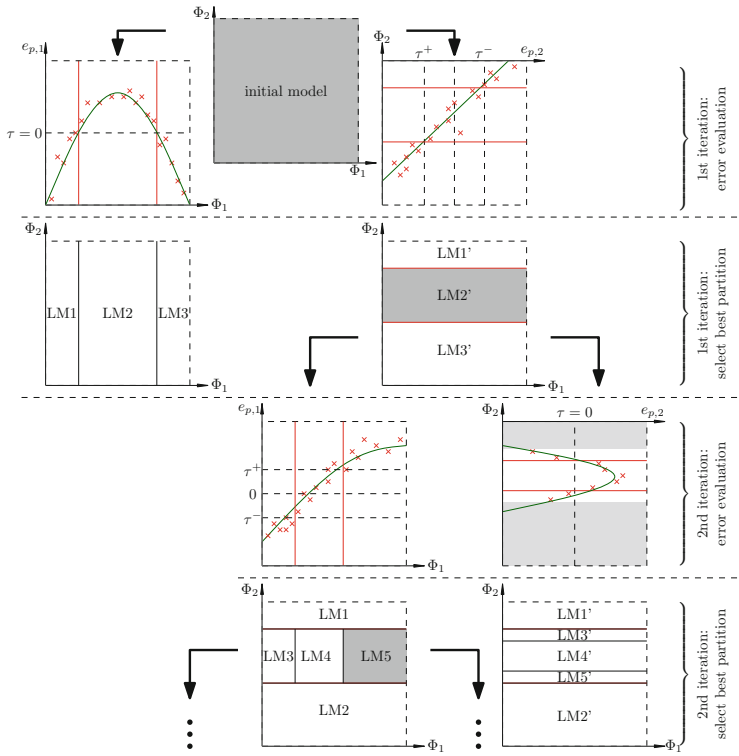


Fig. 5. LoLiMoT algorithm with local error based (LEB) partitioning algorithm.

2. The partitioning criterion is evaluated choosing the worst model. If the chosen model covers the smallest possible region given by the data structure its partitioning criteria is set to zero and the next model is chosen.
3. The local error for the complete operation range is calculated.
4. For every scheduling variable $\Phi_1, \ldots, \Phi_N$ do:
   (a) The projected local error and the partitioning thresholds are calculated. If the original region is minimal regarding the dimension i, the temporary partitioning criterion is set to infinity and the following steps are skipped.
   (b) The model parameters are optimised locally.
   (c) The weighting function and the partitioning criteria of the new models are determined.
5. The best partitioning is chosen and the number of local models increased.
6. If any abort criterion is reached (e.g. maximum number of models), the LMN is completely defined. Otherwise the iteration starts again at point 1.

The inner loop can be evaluated using parallel computing with one threat covering each dimension of the scheduling vector. If linear triangular weighting functions are used, only a part of the weighting functions needs to be updated.

Because every iteration leads to 2 new local models (best case) the approach reduces the number of optimisation tasks from $4N(M-1)/2+1$ optimisation task for the classical LoLiMoT to $3N(M-1)/2+1$ where M denotes the total (odd) number of models and N the dimension of the scheduling vector.

4.2 Results and Comparison of the Partitioning Techniques

We used the LoLiMoT with LEB partitioning technique as well as the classical bisection method to obtain a reduced LMN. To have a fair comparison we ran the latter for more iterations to obtain the same number of local models. Both networks are then compared to the full-grid LMN obtain in Sect. 3 using the verification cycle, see Fig. 6.

Table 1. LMN error development regarding the validation cycle.

	LoLiMoT		Local error based	
M	MSE/hPa^2	$\lvert e_{max}\rvert$/hPa	MSE/hPa^2	$\lvert e_{max}\rvert$/hPa
10	4.1e3	253.9	2.4e3	203.8
30	2.7e3	200.5	2.8e3	165.5
45	2.4e3	208.9	2.3e3	157.5
60	2.4e3	208.9	2.2e3	159.6

Both networks (with only 10 local models each) are able to match the nonlinear dynamics reasonably well. However, the stationary error is significantly larger compared to the full-grid LMN with 256 models. Increasing the number of models both algorithms can reduce the error measure as expected. Often the LEB-network shows better stationary behaviour than the bisection-network, but the data does not allow for a strong statement here. However, the overall error of the LEB-network is significantly better than the errors produced by the bisection-model, see Table 1. This holds in particular for the maximum error which is up to 25 % lower for the LEB-network.

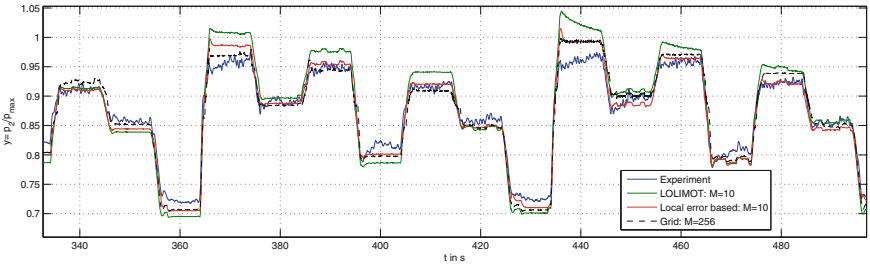


Fig. 6. Comparison of the LMN performances using the verification cycle.

5 Conclusions

In this work we apply a continuous-time global-state local model network to identify the boost-pressure dynamics of an exhaust turbocharger using experimental input-output data. The approach is feasible to match the strongly nonlinear dynamics and may therefore be suitable as a basis for control design. We discuss several implementation issues such as the choice of state-space representation, global-state implementation and the use of discontinuous time-series at various operation points. We propose a novel partition strategy that is computationally more efficient compared to the classical bisection method and also yields better results in the global performance using an independent verification-cycle.

References

1. Murray-Smith, R., Johansen, T.A. (eds.): Multiple Model Approaches to Modelling and Control. Taylor & Francis, London (1997)
2. Nelles, O.: Nonlinear system identification with local linear neuro-fuzzy models. Ph.D. thesis, TU Darmstadt, Aachen (1999)
3. Gawthrop, P.: Continuous-time local state local model networks. In: IEEE International Conference on Systems, Man and Cybernetics, vol. 1, pp. 852–857, Oct 1995
4. Hentabli, K.: State-space local model networks based continuous-time gpc. application to induction motor. In: American Control Conference, vol. 6 (1998)
5. Verdult, V.: Non linear system identification: a state-space approach. Ph.D. thesis, University of Twente, Enschede (2002)
6. Nelles, O.: Axes-oblique partitioning strategies for local model networks. In: Joint CCA, ISIC and CACSD, pp. 2378–2383, Oct 2006
7. Hartmann, B., Skrjan, I., Nelles, O.: Recent partitioning strategies for local model networks. In: Workshop Model-Based Calibration Methods, TU Wien (2011)
8. Johansen, T.A., Foss, B.: Constructing narmax models using armax models. Int. J. Control **58**(5), 1125–1153 (1993)
9. Nelles, O., Sinsel, S., Isermann, R.: Local basis function networks for identification of a turbocharger. In: UKACC International Conference on Control, pp. 7–12 (1996)
10. Nelles, O., Isermann, R.: Basis function networks for interpolation of local linear models. In: 35th IEEE Conference on Decision and Control, pp. 470–475 (1996)
11. Bänfer, O., Hartmann, B., Nelles, O.: Polymot versus hilomot - a comparison of two different training algorithms for local model networks. In: 16th IFAC Symposium on System Identification, pp. 1569–1574 (2012)
12. Hametner, C., Jakubek, S.: Neuro-fuzzy modelling using a logistic discriminant tree. In: American Control Conference, pp. 864–869, July 2007
13. Hartmann, B., Nelles, O.: On the smoothness in local model networks. In: American Control Conference, ACC '09, pp. 3573–3578, Jun 2009
14. Leith, D., Shorten, R., Leithead, W., Mason, O.: Issues in the design of switched linear control systems: a benchmark study. Int. J. Adapt. Control Signal Process. **17**, 103–118 (2003)
15. Wulff, K., Wirth, F., Shorten, R.: A control design method for a class of SISO switched linear systems. Automatica **45**(11), 2592–2596 (2009)

16. Wulff, K., Wirth, F., Shorten, R.: On the stabilisation of a class of SISO switched linear systems. In: 44th IEEE Conference on Decision and Control European Control Conference, CDC-ECC '05, 3976–3981, Dec 2005
17. Moran, M.J., Shapiro, H.N.: Fundamentals of Engineering Thermodynamics, 6th edn. Wiley, Hoboken (2008)
18. Jankovic, M., Jankovic, M., Kolmanovsky, I.: Constructive lyapunov control design for turbocharged diesel engines. Trans. Control Syst. Technol. **8**(2), 288–299 (2000)
19. Jung, M., Glover, K.: Calibratable linear parameter-varying control of a turbocharged diesel engine. Trans. Control Syst. Technol. **14**(1), 45–62 (2006)
20. Schollmeyer, M.: Beitrag zur modellbasierten Ladedruckregelung für Pkw-Dieselmotoren. Ph.D. thesis, Universität Hannover (2010)

Erratum to: More Safe Optimal Input Signals for Parameter Estimation of Linear Systems Described by ODE

Ewaryst Rafajłowicz(✉) and Wojciech Rafajłowicz

Institute of Computer Engineering Control and Robotics, Wroclaw, Poland
ewaryst.rafajlowicz@pwr.wroc.pl

Erratum to:
Chapter 26: C. Pötzsche et al. (Eds.)
System Modeling and Optimization,
DOI:10.1007/978-3-662-45504-3_26

Page 267 - The paper was supported by the National Council for Research of Polish Government under grant 2012/07/B/ST7/01216, internal code 350914 of Wroclaw University of Technology.

Must read:
The paper was supported by The National Science Centre Poland under grant 2012/07/B/ST7/01216, internal code 350914 of Wroclaw University of Technology.

The online version of the original chapter can be found under DOI 10.1007/978-3-662-45504-3_26

C. Pötzsche et al. (Eds.): CSMO 2013, IFIP AICT 443, pp. E1, 2015.
DOI: 10.1007/978-3-662-45504-3_35

Author Index