
ANALYSIS AND OPTIMIZATION OF DIFFERENTIAL SYSTEMS

IFIP – The International Federation for Information Processing

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

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

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

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

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

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

The working conferences are structured differently. They are usually run by a working group and attendance is small and by invitation only. Their purpose is to create an atmosphere conducive to innovation and development. Refereeing is less 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 in information 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.

ANALYSIS AND OPTIMIZATION OF DIFFERENTIAL SYSTEMS

*IFIP TC7 / WG7.2 International Working Conference on
Analysis and Optimization of Differential Systems,
September 10–14, 2002, Constanta, Romania*

Edited by

Viorel Barbu

*University of Iasi
Romania*

Irena Lasiecka

*University of Virginia, Charlottesville
USA*

Dan Tiba

*Weierstrass Institute, Berlin
Germany*

Constantin Varsan

*Institute of Mathematics, Bucharest
Romania*



SPRINGER SCIENCE+BUSINESS MEDIA, LLC

Library of Congress Cataloging-in-Publication Data

A C.I.P. Catalogue record for this book is available from the Library of Congress.

Analysis and Optimization of Differential Systems

Edited by Viorel Barbu, Irena Lasiecka, Dan Tiba, and Constantin Varsan

ISBN 978-1-4757-4506-1 ISBN 978-0-387-35690-7 (eBook)

DOI 10.1007/978-0-387-35690-7

Copyright © 2003 by Springer Science+Business Media New York

Originally published by Kluwer Academic Publishers in 2003

All rights reserved. No part of this work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording, or otherwise, without written permission from the Publisher Springer Science+Business Media, LLC ,
with the exception of any material supplied specifically for the
purpose of being entered and executed on a computer system, for exclusive use by the
purchaser of the work.

Printed on acid-free paper.

Contents

Preface	ix
Compactness and long-time stabilization of solutions to phase-field models <i>Sergiu Aizicovici and Hana Petzeltová</i>	1
PDEs in the inverse problem of dynamics <i>Mira-Cristiana Anisiu</i>	13
Existence and asymptotic behavior for some difference equations associated with accretive operators <i>Narcisa C. Apreutesei</i>	21
Convergence rate of a multiplicative Schwarz method for strongly nonlinear variational inequalities <i>Lori Badea</i>	31
New results about the H -measure of a set <i>Alina Bărbulescu</i>	43
Positional modeling in a system with time delay <i>Marina Blizorukova</i>	49
Uniform stability of a coupled system of hyperbolic/parabolic PDE's with internal dissipation <i>Francesca Bucci</i>	57
Existence of strong solutions of fully nonlinear elliptic equations <i>Adriana Buică</i>	69
Free boundary conditions for intrinsic shell models <i>John Cagnol and Catherine Lebiedzik</i>	77
Error estimates for linear-quadratic elliptic control problems <i>Eduardo Casas and Fredi Tröltzsch</i>	89
Relaxation for quasi-linear differential inclusions in non separable Banach spaces <i>Aurelian Cernea</i>	101

Determining functionals for a class of second order in time evolution equations with applications to von Karman equations <i>Igor Chueshov and Irena Lasiecka</i>	109
On some cases of Cauchy problem <i>Silvia–Otilia Corduneanu</i>	123
Iterative procedure for stabilizing solutions of differential Riccati type equations arising in stochastic control <i>Vasile Dragan, Toader Morozan and Adrian M. Stoica</i>	133
Recursive deconvolution: an overview of some recent results <i>F. Fagnani and L. Pandolfi</i>	145
Determining a semilinear parabolic PDE from final data <i>Luis A. Fernández and Cecilia Pola</i>	157
About a generalized algebraic Riccati equation <i>G. Freiling and A. Hochhaus</i>	169
On the dynamics of contracting relations <i>Vasile Glavan and Valeriu Guțu</i>	179
Level set methods for a parameter identification problem <i>B.–O. Heimsund, T. Chan, T. K. Nilssen and Xue–Cheng Tai</i>	189
Factorization of elliptic boundary value problems: the QR approach <i>Jacques Henry</i>	201
Smooth mappings and non \mathcal{F}_t -adapted solutions associated with Hamilton-Iacobi stochastic equations <i>Daniela Ijacu and Constantin Varsan</i>	211
On the Sobolev boundary value problem with singular and regularized boundary conditions for elliptic equations <i>Nicolae Jitărușu</i>	219
On some optimization problem with non-quadratic criterion <i>Adam Kowalewski</i>	227
Nonconservative Schrödinger equations with unobserved Neumann B.C.: Global uniqueness and observability in one shot <i>I. Lasiecka, R. Triggiani and X. Zhang</i>	235
Optimal flow in dynamic networks with nonlinear cost functions on edges <i>Dmitrii Lozovanu and Dan Stratila</i>	247
Method of extremal shift in problems of reconstruction of an input for parabolic variational inequalities <i>Vyacheslav Maksimov</i>	259

Flow-invariance properties for a class of discrete-time nonlinear uncertain systems <i>Laurentiu Marinovici and Octavian Pastravanu</i>	269
Differential properties of Lipschitz, Hamiltonian and characteristic flows <i>Stefan Mirica</i>	281
The control of Saffman-Taylor instability <i>Gelu Paşa</i>	291
Singular perturbations of hyperbolic-parabolic type <i>Andrei Perjan</i>	297
Improved dynamic properties by feedback for systems with delay in control <i>Dan Popescu and Vladimir Răsvan</i>	303
General connections between strong optimization and Pareto efficiency <i>Vasile Postolică</i>	315
Generalized Ho-Kalman algorithm for 2D continuous discrete linear systems <i>Valeriu Prepeliță</i>	321
Numerical methods for Nash equilibria in multiobjective control of partial differential equations <i>Angel Manuel Ramos</i>	333
Local bifurcation for the FitzHugh-Nagumo system <i>Carmen Rocsoreanu and Mihaela Sterpu</i>	345
Optimization of steady-state flow of incompressible fluids <i>Tomáš Roubiček</i>	357
Topological degree approach to steady state flow <i>Cristina Sburlan and Silviu Sburlan</i>	369
Fast numerical algorithms for Wiener systems identification <i>Vasile Sima</i>	375
Optimization of differential systems with hysteresis <i>Jürgen Sprekels and Dan Tiba</i>	387
Optimal control of non stationary, three dimensional micropolar flows <i>Ruxandra Stavre</i>	399
An H^∞ design method for fault detection and identification problems <i>Adrian M. Stoica and Michael J. Grimble</i>	409

Riccati equation of stochastic control and stochastic uniform observability in infinite dimensions	421
<i>Viorica Mariela Ungureanu</i>	
Componentwise asymptotic stability induced by symmetrical polyhedral time-dependent constraints	433
<i>Mihail Voicu and Octavian Pastravanu</i>	

Preface

The present volume comprises selected contributions presented at the Conference

Analysis and Optimization of Differential Systems

held in Constanța, 10-14 September, 2002, jointly organised by the Institute of Mathematics of the Romanian Academy, Bucharest and the Faculty of Mathematics of the "Ovidius" University in Constanța (Romania).

The Conference is co-sponsored by the Romanian Academy of Sciences and by the Working Group WG 7.2 of IFIP (International Federation for Information Processing), Technical Committee TC7 on "Modeling and Optimization".

This conference was one of the series of international conferences sponsored by WG 7.2 Working Group of IFIP. WG 7.2 conferences are held annually. Most recent events include: Conference on "Modeling and Optimization", Trier, Germany, July 23-26, 2001, Conference on "Optimization of Nonlinear Materials and Structures", Poznan, Poland, August 27-29, 2001, Conference on "Nonlinear Problems in Aviation", Melbourne, Fla, USA, May 15-17, 2002,

The conference was supported by the European Union program EURROMMAT, whose generous financial help is acknowledged and greatly appreciated.

The main theme of the meeting was focused on the qualitative aspects of deterministic and stochastic differential equations. Areas covered by the participants include: control theory of Partial Differential Equations (PDE's); calculus of variations, numerical treatment of solutions to differential equations and related optimization problems; optimization methods in PDE's with numerous applications to mechanics and physics; etc.

Emphasis was placed on topics of high current interest in the research community, where rapid dissemination of information is particularly vital. Most of the presentations dealt with very recent results at the forefront of mathematical interests in their respective disciplines.

The meeting was attended by over 70 participants from 12 countries from Europe, the United States and Canada, in addition to many graduate students and junior researchers. The Central European location of the conference site

allowed for participation of many researchers from Eastern Europe and the former Soviet republics. New scientific contacts and collaborations between these groups of mathematicians have been established during the conference. All talks were well attended and the Conference provided a very lively forum for the exchange of information and for future expansion of the subject matter.

Finally, we wish to thank all the participants for their contributions and for their cooperation in making this workshop a successful event as well as the "Ovidius" University of Constanța for the excellent job of providing high quality service in all logistic aspects of the meeting. Our deep thanks also go to Elena Mocanu from the Institute of Mathematics of the Romanian Academy (Iași branch) who assisted us in the editorial work.

At last, but not at least, we express our gratitude to Kluwer, especially to Yana Lambert, for their efficient and kind cooperation during the preparation of this volume.

It is a pleasure to acknowledge all of these contributions.

THE EDITORS

COMPACTNESS AND LONG-TIME STABILIZATION OF SOLUTIONS TO PHASE-FIELD MODELS

Sergiu Aizicovici

*Department of Mathematics, Ohio University
Morton Hall 321, Athens, OH 45701, U.S.A.*

Hana Petzeltová*

*Institute of Mathematics AV ČR
Žitná 25, 115 67 Praha 1, Czech Republic*

Abstract The compactness of trajectories of solutions to various phase-field models is proved. In some cases, the convergence of any strong solution to a single stationary state is also established.

1. Introduction

The aim of this note is to survey results on convergence of solutions of phase-field models to the stationary states, using a generalization of the Łojasiewicz theorem. We will consider models proposed by Caginalp [5], where the time evolution of the phase variable $\chi(t, x)$ and the temperature $\vartheta(t, x)$ is governed by the system of differential equations:

$$\tau \partial_t \chi = w, \quad (1)$$

$$\partial_t (\vartheta + \lambda(\chi)) + \operatorname{div} \mathbf{q} = 0, \quad (2)$$

where the so called chemical potential w is given by

$$w = \xi^2 \Delta \chi - W'(\chi) + \lambda'(\chi) \vartheta$$

W and λ are given functions, W is typically a double-well potential, and \mathbf{q} denotes the heat flux. We shall also consider the conserved phase-field model, where (1) is replaced by

*The work of this author was supported by Grant A1019002 of GA AV ČR

$$\tau \partial_t \chi = -\xi^2 \Delta w \quad (3)$$

We shall treat the classical case

$$\mathbf{q} = -k_I \nabla \vartheta$$

and also the linearized Coleman–Gurtin [6] constitutive relation, where \mathbf{q} is determined by

$$\mathbf{q} = -k_I \nabla \vartheta - k * \nabla \vartheta, \quad (4)$$

involving the memory effects, where the constant $k_I > 0$ is the instantaneous heat conductivity, k is a suitable dissipative kernel, and the symbol $*$ denotes the time convolution:

$$k * v(t) = \int_0^\infty k(s)v(t-s) \, ds.$$

The material occupies a bounded regular domain $\Omega \subset R^3$ and the system (1)–(2) is complemented by the homogeneous Neumann boundary condition for χ , while ϑ obeys the homogeneous Dirichlet condition.

$$\nabla \chi \cdot \mathbf{n}|_{\partial\Omega} = 0, \quad \vartheta|_{\partial\Omega} = 0. \quad (5)$$

In the conserved model, we require Neumann boundary conditions for both χ , ϑ and also for the chemical potential w which can be expressed by

$$\begin{aligned} \nabla \vartheta \cdot \mathbf{n}|_{\partial\Omega} &= \nabla \chi \cdot \mathbf{n}|_{\partial\Omega} = \nabla(\Delta \chi) \cdot \mathbf{n}|_{\partial\Omega} \\ &\text{with } \mathbf{n} \text{ the outer normal vector.} \end{aligned} \quad (6)$$

For the sake of simplicity, we set the constants τ , ξ , representing a relaxation time and a correlation length, respectively, equal to 1.

Systems of the same or similar type have been recently studied by many authors.(see Colli et al. [7], Giorgi et al. [11], Novick-Cohen [17] etc). The questions of well-posedness and existence of finite dimensional attractors for the conserved model were discussed by Grasselli et al. [12], and the dissipativity of the respective system was studied by Vigni [20]. In particular, the long-time behavior of solutions seems to be well understood and the equilibrium (stationary) solutions have been identified as the only candidates to belong to the ω -limit set of each individual trajectory. If the stationary problem admits only a finite number of solutions, then any solution $\chi(t), \vartheta(t)$ converges as $t \rightarrow \infty$ to a single stationary state. However, the structure of the set of stationary solutions for a general domain may be quite complicated, in particular, the set in question may contain a continuum of nonradial solutions if Ω is a ball or an annulus. If this is the case, it seems highly nontrivial

to decide whether or not the solutions converge to a single stationary state. It is well-known that the convergence of any trajectory might not happen even for finite-dimensional dynamical systems (cf. Aulbach [4]), and similar examples for semilinear parabolic equations were derived by Poláčik and Rybakowski [18]. In 1983, Simon [19] developed a method to study the long-time behaviour of gradient-like dynamical systems based on deep results from the theory of analytic functions of several variables due to Lojasiewicz [16]. Roughly speaking, an analytic function behaves like a polynomial (of a sufficiently high degree) in a neighbourhood of any point where its gradient vanishes (critical points). More specifically, the following assertion holds (see [16, Theorem 4, page 88]):

Proposition 1.1 *Let $G : U(a) \rightarrow C$ be a real analytic function defined on an open neighbourhood $U(a)$ of a point $a \in R^n$. Then there exist $\theta \in (0, \frac{1}{2})$ and $\delta > 0$ such that*

$$|\nabla G(z)| \geq |G(z) - G(a)|^{1-\theta} \text{ for all } z \in R^n, |z - a| < \delta.$$

L. Simon succeeded in proving a generalized version of the above theorem applicable to analytic functionals on Banach spaces. Later on, Jendoubi [15], and Haraux and Jendoubi [13] simplified considerably Simon's original approach making it accessible for application to a broad class of semilinear problems with a variational structure. Related results in this direction were also obtained by Feireisl and Takáč [10], Hoffmann and Rybka [14] etc.

In some cases, Simon's approach can be used to deal with problems with only a partial variational structure. A typical example could be the system (1), (2) with the memory term omitted in (4) (i.e., for $k = 0$). Indeed the "elliptic" part of (1) is the variational derivative of the free energy functional with respect to χ while (2) is not. Since the temperature tends to zero when time is large, it is possible to modify Simon's method to prove convergence of the phase variable χ to a single stationary state, i.e. a solution of the problem

$$\Delta\chi_\infty - W'(\chi_\infty) = 0, \quad \nabla\chi_\infty \cdot \mathbf{n}|_{\partial\Omega} = 0, \quad \vartheta_\infty = 0, \quad (7)$$

under fairly general conditions imposed on λ and W . In the conserved case, the temperature satisfying the Neumann boundary conditions (6) also tends to zero provided that it has zero mean. Considering the model (3), (2), with λ linear and the boundary conditions (6), the quantities $\int_\Omega \chi(t)dx$ and $\int_\Omega \vartheta(t)dx$, are conserved, so we can normalize the initial functions $\chi(0), \vartheta(0)$ such that $\int_\Omega \chi(0)dx = 0, \int_\Omega \vartheta(0)dx = 0$ which leads to the convergence $\chi \rightarrow \chi_\infty, \vartheta \rightarrow \vartheta_\infty$, where

$$\Delta(\Delta\chi_\infty - W'(\chi_\infty)) = 0, \quad \nabla\chi_\infty \cdot \mathbf{n}|_{\partial\Omega} = \nabla(\Delta\chi_\infty) \cdot \mathbf{n}|_{\partial\Omega} = 0, \quad \vartheta_\infty = 0, \quad (8)$$

or $\vartheta_\infty = \text{const}$ in the general case.

Using the *summed past history* of ϑ , introduced by Dafermos, we can obtain similar results when the memory effects are taken into account in (4), both for the conserved and non-conserved models.

We conclude this contribution with *a priori* estimates that imply compactness of trajectories of solutions of the most general conserved model with memory term and a nonlinear λ . These estimates can be used to prove the existence of strong global solutions emanating from sufficiently smooth data as well as the convergence of solutions satisfying the Poincaré inequality. The existence of global weak solutions vanishing on a nontrivial (positive measure) part of the boundary of Ω and, therefore, satisfying the Poincaré inequality is stated in [12].

2. Main results

In this section, we present a synthesis of some convergence results from the papers [1], [2], [3], [8].

Theorem 2.1 *Let $\Omega \subset R^3$ be a bounded domain of class $C^{2+\mu}$, $\mu > 0$. Suppose, moreover, that the nonlinearities λ, W satisfy the following hypotheses:*

The function λ is of class $C^{1+\mu}(R)$, $\lambda(0)=0$, $|\lambda'(z)| \leq \Lambda$, $z \in R$; (9)
The "free energy" function W is real analytic on R .

In addition, we assume that the instantaneous heat conductivity $k_I > 0$ is strictly positive and the kernel k satisfies:

$$\begin{aligned} k &\in L^1(0, \infty), \quad k \text{ is convex on } (0, \infty), \\ dk'(s) + \delta k'(s) \, ds &\geq 0 \text{ for a certain } \delta > 0. \end{aligned} \quad (10)$$

Let χ, ϑ be a globally defined strong solution of the problem (1), (2), (4), (5) such that

$$\sup_{t>0} \left(\sup_{x \in \Omega} (|\chi(t, x)| + |\vartheta(t, x)|) \right) < \infty. \quad (11)$$

Then there exists χ_∞ - a solution of the stationary problem (7) such that

$$\chi(t) \rightarrow \chi_\infty, \quad \vartheta(t) \rightarrow 0 \text{ in } C(\bar{\Omega}) \text{ as } t \rightarrow \infty.$$

Theorem 2.2 *Let the assumptions of the Theorem 1 be satisfied and*

$$\lambda(z) = cz \quad \text{for some real } c.$$

Let χ, ϑ be a globally defined strong solution of the problem (3), (2), (4), (6) satisfying (11). Then there exists χ_∞ - a solution of the stationary problem (8) such that

$$\chi(t) \rightarrow \chi_\infty, \vartheta(t) \rightarrow \text{const in } C(\bar{\Omega}) \text{ as } t \rightarrow \infty.$$

In the proof, using the fact that the integral means $\int_{\Omega} \chi(t, x) dx$, $\int_{\Omega} \vartheta(t, x) dx$ are conserved quantities, we normalize the initial functions $\chi(0)$, $\vartheta(0)$ to be of zero mean and work in the corresponding spaces, where the solution of the problem

$$-\Delta v = g \text{ (in) } \Omega, \quad \nabla v \cdot \mathbf{n} = 0 \text{ on } \partial\Omega, \quad \int_{\Omega} v dx = 0$$

is uniquely defined and denoted by $v = -\Delta_N^{-1}[g]$. In this space, the Poincaré inequality takes place. See [8] for the proof of Theorem 2.2 when $k = 0$ and [3] when the memory is included.

Remark 1. Here a globally defined strong solution means that $\chi_t, \vartheta_t, D_x^k \chi, D_x^2 \vartheta$ are in the space $L_{loc}^r(0, \infty; L^2(\Omega))$ for any $r \geq 1$, and the boundary conditions are satisfied for all $t \in (0, \infty)$. Moreover, $\chi(0)$ is supposed to belong to $W^{2,2}(\Omega)$ and the past values of ϑ are given for $t \in (-\infty, 0]$, and $\|\vartheta(t)\|_{W^{2,2}(\Omega)}$ are bounded uniformly for $t \in (-\infty, 0]$ and satisfy the boundary conditions (6) when we treat the conserved problem with memory.

Remark 2. A typical example of a kernel k satisfying (10) is $k(s) = s^{-\alpha} e^{-\beta s}$, $0 \leq \alpha < 1$, $\beta > 0$.

Remark 3. The assumption that χ, ϑ is a strong solution of the problem is not restrictive. It will be clear from the estimates presented in Section 3 that any weak solution emanating from smooth initial data will be globally defined and regular on the interval $(0, \infty)$. Moreover, those estimates also allow for more general energy functionals W than the ones considered in Grasselli, Pata and Vugni [12] and Vugni [20].

Sketch of proofs.

First, we derive necessary *a priori* estimates to show that the trajectories $\cup_{t \geq 1} \vartheta(t)$, $\cup_{t \geq 1} \chi(t)$ are precompact in $C(\bar{\Omega})$ and $\vartheta(t) \rightarrow 0$. From the strong maximum principle we deduce that the ω -limit set $\omega[\chi]$ is contained in some interval $[-L, L]$. Accordingly, since we are interested in the ω -limit set of one particular trajectory which is uniformly bounded with respect to χ -component, we are allowed to suppose, without loss of generality, that W' has been modified outside of the interval $[-2L, 2L]$ in such a way that

$$W'(z) \text{ is real analytic on } (-L, L); \tag{12}$$

$$|W''(z)|, |W'(z)| \text{ are uniformly bounded for } z \in R. \quad (13)$$

The next step is to show that $\chi_t \in L_1(T, \infty; X)$ where X denotes a suitable space. To this end, we apply Simon's method to the functional

$$I(v) = \int_{\Omega} (|\nabla v|^2 + W(v)) \, dx \quad (14)$$

to obtain the following generalization of Proposition 1.1.

Proposition 2.1 *Let W satisfy the hypotheses (12), (13). Let $w \in W_N^{2,p}$,*

$$-L < w(x) < L \text{ for all } x \in \Omega.$$

Then for any $P > 0$ there exist constants $\theta \in (0, 1/2)$, $M(P)$, $\varepsilon(P)$ such that

$$|I(v) - I(w)|^{1-\theta} \leq M \| -\Delta v + W'(v) \|_{[W_N^{1,2}(\Omega)]^*} \quad (15)$$

holds for any $v \in W_N^{1,2}(\Omega)$ such that

$$\|v - w\|_{L^2(\Omega)} < \varepsilon, \quad |I(v) - I(w)| < P. \quad (16)$$

The proof is identical with [9, Section 6, Proposition 6.1].

The energy equality, obtained by multiplying the equation (1) by χ_t , (2) by ϑ , ((3) by $-\Delta_N^{-1}[\chi_t]$ respectively), integrating the resulting expressions by parts and adding up, reads:

$$\begin{aligned} \frac{d}{dt} \left[\int_{\Omega} \left(\frac{1}{2} |\nabla \chi(t)|^2 + \frac{1}{2} |\vartheta(t)|^2 + W(\chi(t)) \right) dx + \frac{1}{2} \int_0^{\infty} (-k')(s) \|\nabla \eta(t, s)\|_{L^2(\Omega)}^2 ds \right] \\ + \left\| \left(-\Delta \right)_N^{-\frac{1}{2}} [\chi_t(t)] \right\|_{L^2(\Omega)}^2 + k_I \|\nabla \vartheta(t)\|_{L^2(\Omega)}^2 \\ + \frac{1}{2} \int_0^{\infty} \|\nabla \eta(t, s)\|_{L^2(\Omega)}^2 dk'(s) = 0. \end{aligned} \quad (17)$$

Here we used the summed past history of ϑ defined by

$$\eta(t, s, x) = \int_{t-s}^t \vartheta(z, x) dz, \quad s \geq 0,$$

and the relation

$$\begin{aligned} \int_{\Omega} (k * \vartheta) \vartheta \, dx = \\ \frac{1}{2} \left[\frac{d}{dt} \int_0^{\infty} (-k')(s) \|\eta(t, s)\|_{L^2(\Omega)}^2 ds + \int_0^{\infty} (-k)'(s) \frac{\partial}{\partial s} \|\eta(t, s)\|_{L^2(\Omega)}^2 ds \right]. \end{aligned} \quad (18)$$

Denoting by E the "total energy",

$$E(t) = \frac{1}{2} \int_{\Omega} |\nabla \chi(t)|^2 + 2W(\chi(t)) \, dx + \frac{1}{2} \|\vartheta\|_{L^2(\Omega)}^2 + \frac{1}{2} \int_0^\infty (-k')(s) \|\nabla \eta(t, s)\|_{L^2(\Omega)}^2 \, ds,$$

and taking into account (10), we have $E(t) \rightarrow E_\infty$ as $t \rightarrow \infty$. Moreover, we can prove that

$$\|\vartheta(t)\|_{L^2(\Omega)} \rightarrow 0 \text{ as } t \rightarrow \infty, \quad (19)$$

and

$$\int_0^\infty (-k')(s) \|\nabla \eta(t, s)\|_{L^2(\Omega)}^2 \, ds \rightarrow 0 \text{ as } t \rightarrow \infty. \quad (20)$$

We have

$$I(\chi(t)) \rightarrow I_\infty = E_\infty = \frac{1}{2} \int_{\Omega} |\nabla \chi_\infty|^2 + 2W(\chi_\infty) \, dx \text{ for any } \chi_\infty \in \omega[\chi].$$

In particular, the energy of solutions $\chi_\infty \in \omega[\chi]$ equals the same constant I_∞ .

We integrate (17) with respect to t , and make use of Proposition 2.1 to conclude that there exists $T > 0$ such that

$$\int_T^\infty \|\chi_t(t)\|_X \, dt < \infty,$$

which, together with the compactness of trajectories yields the assertions of Theorems 2.1 and 2.2.

3. A priori estimates

In this section, we prove *a priori* estimates for solutions of the problem (3), (2), (4) with a nonlinear function λ satisfying

$$\lambda \in C^3(R), \quad |\lambda'(r)| \leq \Lambda, \quad |\lambda''(r)| \leq \Lambda \text{ for a certain } \Lambda > 0 \quad (21)$$

The free energy functional $W : R \mapsto R$ will be supposed to satisfy the following hypotheses:

- $W(z) \geq 0$ for all $z \geq 0$, (22)
- $W'(z)z > 0$ for $|z| > 1$, (23)
- $W'(z)z \geq c_1 W(z) - c_2$ for all $z \in R$, (24)

$$W''(z) \geq -c_3, \quad (25)$$

$$W \in C^{3+\mu}(R), |W'(z)| \leq c_4(1 + |z|^{p-1}), p < 5. \quad (26)$$

The energy estimate (17) gives

Lemma 3.1 *Under the hypotheses of Theorem 2.2, there exists E_0 depending only on the quantities*

$$\sup_{t \in (-\infty, 0]} \|\nabla \vartheta(t)\|_{L^2(\Omega)}, \|\nabla \chi(0)\|_{L^2(\Omega)}, \|\chi(0)\|_{L^\infty(\Omega)}$$

such that

$$\sup_{t > 0} \|\vartheta(t)\|_{L^2(\Omega)} + \sup_{t > 0} \|\nabla \chi(t)\|_{L^2(\Omega)} \leq E_0, \quad (27)$$

$$\int_0^\infty \|(-\Delta_N)^{-\frac{1}{2}}[\chi_t(t)]\|_{L^2(\Omega)}^2 + \|\nabla \vartheta(t)\|^2 s dt \leq E_0 \quad (28)$$

Next, we multiply (3) by χ to deduce

$$\frac{d}{dt} \frac{1}{2} \|\chi\|_{L^2(\Omega)}^2 + \|\Delta \chi\|_{L^2(\Omega)}^2 + \int_\Omega W''(\chi) |\nabla \chi|^2 dx = - \int_\Omega \lambda'(\chi) \vartheta \Delta \chi dx,$$

Consequently, (27), (28), (25) and the Poincaré and Young inequalities imply

$$\int_t^{t+1} \|\Delta \chi\|_{L^2(\Omega)}^2 d\tau \leq E_0 \text{ for any } t \geq 0, \quad (29)$$

To improve the estimates on χ , we write (3) as an evolutionary equation

$$\frac{\partial \chi}{\partial t} + \Delta^2 \chi = \Delta[W'(\chi)] - \Delta[\lambda'(\chi) \vartheta]. \quad (30)$$

Let p be as in (26). We prove first that

$$\chi \in L^r(t, t+1; W^{2,q_1}(\Omega)), t \geq 0, \text{ for any } 1 \leq r < \infty, q_1 = \min\{2, 6/p\}.$$

For this, for all $1 < q < \infty$, we define a linear operator $\Delta_{N,q}$ on the Banach space $L^q(\Omega)$ by

$$\mathcal{D}(\Delta_{N,q}) = \{v \in W^{2,q}(\Omega) \mid \nabla v \cdot \vec{n} = 0 \text{ on } \partial\Omega\}, \Delta_N v = \Delta v,$$

and rewrite (30) in the abstract form

$$\chi_t + \Delta_{N,q}^2.$$

From (27), (21), we know that h_2 is bounded in $L^\infty(t, t+1; [W^{2,2}(\Omega)]^*)$ uniformly for all $t \geq 0$. On the other hand, using (27) and the Sobolev embedding $W^{1,2}(\Omega) \subset L^6(\Omega)$, we have $\chi \in L^\infty(0, \tau; L^6(\Omega))$ for all $\tau > 0$. From (26) we get $W'(\chi) \in L^\infty(0, \tau; L^{\frac{6}{p}}(\Omega))$.

Also, $\Delta_{N,q}^{-1}(\Delta_{N,q})f = f - \int_\Omega f(x)dx$ for $f \in L^q(\Omega)$. Hence

$$\begin{aligned} \|h\|_{\mathcal{D}(\Delta_{N,q}^{-1})} &= \| [W'(\chi) - \lambda'(\chi)\vartheta] - \frac{1}{|\Omega|} \int_\Omega [W'(\chi) - \lambda'(\chi)\vartheta] dx \|_{L^q(\Omega)} \\ &\leq C \left(\|W'(\chi)\|_{L^q(\Omega)} + \|\vartheta\|_{L^q(\Omega)} \right). \end{aligned}$$

This implies that $\chi \in L^r(t, t+1; W^{2,q_1}(\Omega))$ where $q_1 = \min\{2, \frac{6}{p}\}$, $r \geq 1$. Consequently, by the Sobolev embedding theorem.

$$\chi \in L^r(t, t+1; L^{q_2}(\Omega)) \text{ with } q_2 = \frac{3q_1}{3 - 2q_1} \text{ if } 2q_1 < 3, q_2 = \infty \text{ otherwise.}$$

Next we argue by induction (bootstrap argument). We deduce from (26) that

$$W'(\chi) \in L^{\frac{r}{p}}(t, t+1; L^{\frac{q_2}{p}}(\Omega)).$$

Remark that we have

$$\frac{q_2}{p} - q_1 \geq \frac{6}{p(p-4)} - \frac{6}{p} > 0$$

provided $p \in (4, 5)$, $q_2 = \infty$ if $p \leq 4$. Hence, after a finite number of steps we arrive at the estimate

$$\chi \in L^r(t, t+1; W^{2,2}(\Omega)) \subset L^r(t, t+1; L^\infty(\Omega)), \quad t \geq 0 \quad (31)$$

for any $1 \leq r < \infty$. Also, $\chi_t \in L^r(t, t+1; [W^{2,2}(\Omega)]^*)$ which implies

$$\chi \in C\left(t, t+1; (W^{2,2}(\Omega), [W^{2,2}(\Omega)]^*)_\theta\right),$$

with θ satisfying $\theta(1 - \frac{1}{r}) > \frac{1-\theta}{r}$, where $(X, Y)_\theta$ denotes the interpolation space. As $r > 1$ is arbitrary, we can choose θ such that $(W^{2,2}(\Omega), [W^{2,2}(\Omega)]^*)_\theta \hookrightarrow C(\overline{\Omega})$.

$$\sup_{t>0} \|\chi(t)\|_{C(\overline{\Omega})} \leq C_\infty. \quad (32)$$

This implies that $W''(\chi)$, $W'''(\chi)$ are bounded, and $\nabla\chi$ is bounded in $L^r(t, t+1; L^6(\Omega))$ for all r , independently of $t > 0$. Then

$$\int_t^{t+1} \|\Delta W'(\chi(s))\|_{L^2(\Omega)}^2 ds < C \text{ for all } t > 0. \quad (33)$$

Moreover, by (27), (28),

$$\vartheta \in L^\infty(t, t+1; L^2(\Omega)) \cap L^2(t, t+1; W^{1,2}(\Omega)) \hookrightarrow L^s(t, t+1; L^3(\Omega)) \text{ for } s < 4.$$

By (31),

$$\nabla \chi \in L^r(t, t+1; L^6(\Omega)) \text{ for all } 1 \leq r < \infty. \quad (34)$$

It follows that $\nabla(\lambda'(\chi)\vartheta) \in L^2(t, t+1; L^2(\Omega))$ and, by the same reasoning as above

$$\chi \in L^2(t, t+1; W^{3,2}(\Omega)), \quad t \geq 0. \quad (35)$$

This yields

$$\lambda(\chi) \in L^2(t, t+1; W^{3,2}(\Omega)), \quad t \geq 0. \quad (36)$$

In fact, $\lambda'''(\chi)$ is bounded because of (32) and

$$\Delta \chi \in L^2(t, t+1; L^6(\Omega)) \cap L^r(t, t+1; L^2(\Omega)) \Rightarrow \Delta \chi \in L^s(t, t+1; L^3(\Omega)), \quad s < 4,$$

which, together with (34) gives

$$|\nabla \chi \cdot \Delta \chi| \in L^2(t, t+1; L^2(\Omega)), \quad |\nabla \chi|^3 \in L^r(t, t+1; L^2(\Omega)).$$

Now, we multiply (2) by $\Delta(\vartheta + \lambda(\chi))$, integrate by parts and use (18) with $\Delta \vartheta$ in place of ϑ , to obtain

$$\begin{aligned} & \frac{1}{2} \frac{d}{dt} \left[\|\nabla(\vartheta + \lambda(\chi))\|_{L^2(\Omega)}^2 + \int_0^\infty (-k')(s) \|\Delta \eta(t, s)\|_{L^2(\Omega)}^2 ds \right] \\ & + \|\Delta \vartheta\|_{L^2(\Omega)}^2 + \int_\Omega \Delta \vartheta \Delta \lambda(\chi) dx + \int_0^\infty \|\Delta \eta(t, s)\|_{L^2(\Omega)}^2 dk'(s) = \\ & \quad \int_\Omega k * \nabla \vartheta \cdot \nabla \Delta \lambda(\chi) dx. \end{aligned}$$

If we denote

$$F(t) = \|\nabla(\vartheta + \lambda(\chi))\|_{L^2(\Omega)}^2 + \int_0^\infty (-k')(s) \|\Delta \eta(t, s)\|_{L^2(\Omega)}^2 ds,$$

employ (10), the Poincaré and Young inequalities, we obtain

$$\frac{d}{dt} F(t) + aF(t) \leq C \left(1 + \|\lambda(\chi(t))\|_{W^{3,2}(\Omega)}^2 + k * \|\nabla \vartheta\|_{L^2(\Omega)}^2(t) \right),$$

for some small $a > 0$, which yields the estimate

$$F(t) \leq C \left(1 + \sup_{t>0} \int_t^{t+1} \|\lambda(\chi(s))\|_{W^{3,2}(\Omega)}^2 + k * \|\nabla \vartheta\|_{L^2(\Omega)}^2(s) ds \right). \quad (37)$$

We thereby arrive at

Lemma 3.2 *Under the hypotheses of Theorem 2.2, there exists E_0 depending only on the quantities $\sup_{t \in (-\infty, 0]} \|\Delta\vartheta(t)\|_{L^2(\Omega)}$, $\|\Delta\chi(0)\|_{L^2(\Omega)}$, such that*

$$\sup_{t \geq 0} \|\nabla\vartheta(t)\|_{L^2(\Omega)} \leq E_0, \quad (38)$$

$$\int_t^{t+1} \|\Delta\vartheta(s)\|^2 ds \leq E_0 \text{ for all } t \geq 0. \quad (39)$$

We continue the bootstrapping, and use Lemma 3.2 together with (31) and (35) to deduce

$$\Delta[\lambda'(\chi)\vartheta] \in L^2(t, t+1; L^2(\Omega)), \quad t \geq 0. \quad (40)$$

By virtue of (33), (40), the phase field variable χ satisfies the equation (30) with the right hand side bounded in $L^2(t, t+1; L^2(\Omega))$ independently of t . Therefore, we obtain

$$\chi \in L^2(t, t+1; W^{4,2}(\Omega)), \quad \chi_t \in L^2(t, t+1; L^2(\Omega)), \quad t > 1. \quad (41)$$

The convolution term in (2) is bounded in $L^2(\Omega)$ according to (10), (39), and $\Delta\lambda(\chi)$ is bounded in $L^r(t, t+1; L^2(\Omega))$ for any r independently of $t > 0$ by (31), (33). Hence the equation (2) can be written as a parabolic equation for $e = \vartheta + \lambda(\chi)$

$$e_t - \Delta e = -\Delta\lambda(\chi) + k * \Delta\vartheta$$

with the right-hand side bounded in $L^r(t, t+1; L^2(\Omega))$ for any $r \geq 1$ and any $t > 0$. This gives $\Delta\vartheta \in L^r(t, t+1; L^2(\Omega))$, and, consequently

$$\|\chi_t\|_{L^r(t, t+1; L^2(\Omega))} \leq C, \quad \text{for any } r \geq 1, \quad t > 0, \quad \text{and then also}$$

$$\|\vartheta_t\|_{L^r(t, t+1; L^2(\Omega))} \leq C, \quad \text{for any } r \geq 1, \quad t > 0$$

In particular we have obtained the following result:

Proposition 3.1 *Let $\Omega \subset R^N$, $N \leq 3$ be a bounded domain with boundary of class $C^{2+\mu}$, $\mu > 0$. Let $\lambda, W \in C^{3+\mu}(R)$ satisfy hypotheses (21)–(26). Then for any strong solution χ, ϑ of the problem (3), (2), (4), (6) on the time interval $(0, \infty)$, the trajectories $\bigcup_{t \geq 1} \chi(t)$, $\bigcup_{t \geq 1} \vartheta(t)$ are precompact in the space $C(\overline{\Omega})$. Moreover, $\bigcup_{t \geq 1} \chi(t)$ is precompact in $W^{1,2}(\Omega)$.*

References

- [1] S. Aizicovici and E. Feireisl. Long-time stabilization of solutions to a phase-field model with memory. *J. Evolution Equations.*, **1**:69–84, 2001.
- [2] S. Aizicovici, E. Feireisl, and F. Issard-Roch. Long-time convergence of solutions to the phase-field system. *Math. Meth. Appl. Sci.*, **24**:277–288, 2001.
- [3] S. Aizicovici and H. Petzeltová. Asymptotic behavior of solutions of a conserved phase-field system with memory. *J. Integral Equations Appl.* submitted.
- [4] B. Aulbach. *Continuous and discrete dynamics near manifolds of equilibria*. Lecture Notes in Math. 1058, Springer-Verlag, New York, 1984.
- [5] G. Caginalp. An analysis of a phase field model of a free boundary. *Arch. Rational Mech. Anal.*, **92**:205–246, 1986.
- [6] B.D. Coleman and M.E. Gurtin. Equipresence and constitutive equations for rigid heat conductors. *Z. Angew. Math. Phys.*, **18**:199–208, 1967.
- [7] P. Colli, G. Gilardi, P. Laurençot, and A. Novick-Cohen. Uniqueness and long-time behaviour for the conserved phase-field system with memory. *Discrete Contin. Dynam. Systems*, **5**:375–390, 1999.
- [8] E. Feireisl, F. Issard-Roch, and H. Petzeltová. On the long-time behaviour and convergence towards equilibria for a conserved phase-field model. *Discrete Contin. Dynam. Systems*, 2002. To appear.
- [9] E. Feireisl and F. Simondon. Convergence for semilinear degenerate parabolic equations in several space dimensions. *J. Dynamics Differential Equations*, **12**(3):647–673, 2000.
- [10] E. Feireisl and P. Takáč. Long-time stabilization of solutions to the Ginzburg-Landau equations of superconductivity. *Monatsh. Math.*, **133**:197–221, 2001.
- [11] C. Giorgi, M. Grasselli, and V. Pata. Well-posedness and longtime behavior of the phase-field model with memory in a history space setting. *Quart. Appl. Math.* 2000. To appear.
- [12] M. Graselli, V. Pata, and F.M. Vegni. Longterm dynamics of a conserved phase-field system with memory. 2001. Preprint.
- [13] A. Haraux and M.A. Jendoubi. Convergence of solutions of second-order gradient-like systems with analytic nonlinearities. *J. Differential Equations*, **144**:313–320, 1998.
- [14] K.-H. Hoffmann and P. Rybka. Convergence of solutions to Cahn-Hilliard equation. *Commun. Partial Differential Equations*, **24**:1055–1077, 1999.
- [15] M. A. Jendoubi. Convergence of global and bounded solutions of the wave equation with linear dissipation and analytic nonlinearity. *J. Differential Equations*, **144**(2):302–312, 1998.
- [16] S. Łojasiewicz. Une propriété topologique des sous ensembles analytiques réels. *Colloques du CNRS, Les équations aux dérivées partielles*, **117**, 1963.
- [17] A. Novick-Cohen. Conserved phase-field equations with memory. In *Curvature flows and related topics (Levico 1994)*, pages 179–197. GAKUTO Internat. Ser. Math. Sci. Appl. 5, 1995, Tokyo.
- [18] P. Poláčik and K. P. Rybakowski. Nonconvergent bounded trajectories in semilinear heat equations. *J. Differential Equations*, **124**:472–494, 1996.
- [19] L. Simon. Asymptotics for a class of non-linear evolution equations, with applications to geometric problems. *Ann. of Math.*, **118**:525–571, 1983.
- [20] F.M. Vegni. Dissipativity of a conserved phase-field model with memory. 2001, Preprint.

PDES IN THE INVERSE PROBLEM OF DYNAMICS

Mira-Cristiana Anisiu

T. Popoviciu Institute of Numerical Analysis

Romanian Academy

37 Republicii st., Cluj-Napoca, RO-3400 Romania

mira@math.ubbcluj.ro

Abstract The basic equations are exposed for the following version of the inverse problem of dynamics: determine the two-dimensional potential compatible with a given family of orbits, traced by a material point. If the potential is known in advance, a nonlinear equation is satisfied by the function representing the family of orbits. Its solutions are studied in the presence of additional information on the family. The possibility of programming the motion of a material point in a preassigned region of the plane is also considered.

Keywords: inverse problem of dynamics, Szebehely's and Bozis' equations.

1. Introduction

The version of the inverse problem of dynamics discussed in this paper consists in finding the two-dimensional potential which governs the motion of a dynamical system, knowing a given family of orbits. The first outstanding results are due to Newton (1687), who found the force law compatible with Kepler's laws. The paper which gave a new impulse to the field of the inverse problems was that of Szebehely's (1974). Information on various other aspects (nonconservative systems, three-dimensional ones, rotating frames, holonomic systems with n degrees of freedom etc) of the inverse problem is contained in Bozis (1995) and Anisiu (1998).

2. The main tools of the inverse problem

Given a family of curves

$$f(x, y) = c \quad (2.1)$$

one looks for a potential for which this family is an orbit family of a particle. The problem is then to determine the potential $V \in C^1(D)$ for which the equations of the motion of a unit mass particle are

$$\ddot{x} = -V_x \quad \ddot{y} = -V_y, \quad (2.2)$$

knowing a given family of orbits (2.1). The first order partial differential equation satisfied by the potential V was obtained by Szebehely (1974).

The system (2.2) admits the energy integral $\dot{x}^2 + \dot{y}^2 = 2(E(f) - V)$, $E(f)$ being constant on every orbit in (2.1).

Theorem 2.1 *Let $D \subset \mathbb{R}^2$ be an open set and $f \in C^2(D)$ such that $f_x^2(x, y) + f_y^2(x, y) \neq 0$ for each $(x, y) \in D$. If the system (2.2) admits as orbits the curves of the family (2.1), then the function V satisfies the partial differential equation*

$$f_x V_x + f_y V_y - \frac{2(E(f) - V)}{f_x^2 + f_y^2} (f_{xx} f_y^2 - 2f_{xy} f_x f_y + f_{yy} f_x^2) = 0. \quad (2.3)$$

Proof. From (2.1) we obtain $\dot{x} f_x + \dot{y} f_y = 0$, hence \dot{x} and \dot{y} will be

$$\dot{x} = \pm \sqrt{k} f_y \quad \dot{y} = \mp \sqrt{k} f_x, \quad (2.4)$$

where the sign depends on the sense of the motion on the orbit, and $k \in C^1(D, \mathbb{R}_+)$ is an arbitrary function. Differentiating again we have

$$\begin{aligned} \ddot{x} &= k (f_{xy} f_y - f_{yy} f_x) + (k_x f_y - k_y f_x) f_y / 2 \\ \ddot{y} &= k (f_{xy} f_x - f_{xx} f_y) - (k_x f_y - k_y f_x) f_x / 2. \end{aligned}$$

It follows

$$\begin{aligned} -V_x &= k (f_{xy} f_y - f_{yy} f_x) + (k_x f_y - k_y f_x) f_y / 2 \\ -V_y &= k (f_{xy} f_x - f_{xx} f_y) - (k_x f_y - k_y f_x) f_x / 2 \end{aligned}$$

and

$$f_x V_x + f_y V_y = -k (2f_{xy} f_x f_y - f_x^2 f_{yy} - f_y^2 f_{xx}). \quad (2.5)$$

From the energy integral and the relations (2.4), the function k is given by $k = 2(V - E(f)) / (f_x^2 + f_y^2)$. We replace it in (2.5) and obtain Szebehely's equation (2.3). ■

Remark 2.1 In the paper Bozis (1983) the functions

$$\gamma = \frac{f_y}{f_x} \quad \Gamma = \gamma \gamma_x - \gamma_y \quad (2.6)$$

were introduced for the family of curves (2.1). Using these notations, Szebehely's equation was written in the simpler form (Bozis, 1983)

$$V_x + \gamma V_y + \frac{2\Gamma(E - V)}{1 + \gamma^2} = 0. \quad (2.7)$$

From the equation above, an inequality can be deduced because the kinetic energy of the particle $B = E - V$ is nonnegative.

Corollary 2.1 *The potential V satisfies the inequality (Bozis and Ichtiaroglou, 1994)*

$$\Gamma(V_x + \gamma V_y) \leq 0. \quad (2.8)$$

In what follows the functions are considered to be sufficiently smooth.

The equation (2.7) has the disadvantage that it contains the energy E , which in general is not known in advance. Bozis (1984) has eliminated the energy, obtaining a second order partial differential equation.

Theorem 2.2 *Denoting $\kappa = \frac{1}{\gamma} - \gamma$, $\lambda = \frac{\Gamma_y - \gamma\Gamma_x}{\gamma\Gamma}$, $\mu = \lambda\gamma + \frac{3\Gamma}{\gamma}$, the potential V satisfies the second order equation (Bozis, 1984)*

$$-V_{xx} + \kappa V_{xy} + V_{yy} = \lambda V_x + \mu V_y. \quad (2.9)$$

Proof. The energy $E = E(f)$ can be eliminated from equation (2.7) by solving it with respect to E and inserting E into

$$E_y = \gamma E_x, \quad (2.10)$$

the slope function γ being given by (2.6). Relation (2.10) holds because from $E = E(f)$ we have $E_x = E' f_x$ and $E_y = E' f_y$. ■

The main tools used in the inverse problem are: Szebehely's quasi-linear first order partial differential equation (2.7), Bozis' linear second order partial differential equation (2.9), and the inequality (2.8). If one manages to obtain a solution of (2.9), from (2.7) one can find the value of the energy, and from (2.8) the plane region where the actual trajectories are allowed to take place.

Example 2.1 For the case of the family $f(x, y) = x^2 + y^2$ of concentric circles, Broucke and Lass (1977) found (in polar coordinates r, θ) the potential $V(r, \theta) = g(r) + h(\theta)/r^2$, where g and h are arbitrary functions. The energy function is in this case $E = g(r) + rg'(r)/2$, and the allowed region obtained from (2.8) is given by $g'(r) - 2h(\theta)/r^3 \geq 0$. Other cases in which Szebehely's equation is solvable are exposed in Grigoriadou et al (1999).

The relation (2.9) was derived only for the case $\Gamma \neq 0$. From (2.6) one can express Γ in terms of the derivatives of f as

$$\Gamma = \frac{2f_{xy}f_xf_y - f_{xx}f_y^2 - f_{yy}f_x^2}{f_x^3},$$

hence the family (2.1) consists of straight lines if and only if $\Gamma = 0$; in view of (2.6) this condition may be written as

$$\gamma\gamma_x - \gamma_y = 0. \quad (2.11)$$

3. Families of straight lines

The problem of determining the potentials under whose action families of straight lines are described was considered by Bozis and Anisiu (2001).

Szebehely's equation (2.7) for a family of straight lines becomes

$$V_x + \gamma V_y = 0. \quad (3.12)$$

Remark 3.1 In what follows we shall consider $\gamma \neq 0$, because for $\gamma = 0$ we obtain $V_x = 0$, hence $V = v(y)$; this potential produces only the family of vertical straight lines $x = c$.

We differentiate (3.12) with respect to x , then to y , and express γ_x and γ_y from the obtained equations. From (2.11) we get

$$-V_{xx} + \left(\frac{1}{\gamma} - \gamma\right)V_{xy} + V_{yy} = 0; \quad (3.13)$$

this corresponds to Bozis' partial differential equation for the straight lines situation. We insert γ obtained from (3.12) into (3.13) and get a nonlinear partial differential equation

$$V_x V_y (V_{xx} - V_{yy}) = V_{xy} (V_x^2 - V_y^2) \quad (3.14)$$

which must be satisfied by all potentials creating (among other orbits) a family of straight lines.

Remark 3.2 Given an adequate γ (i.e. satisfying (2.11)), there exist infinitely many potentials $V(x, y)$, found from (3.12), creating the monoparametric family of straight lines. But for an adequate V (i.e. satisfying (3.14) and depending on both variables x and y), there corresponds to it exactly one γ , found from (3.12), hence one family of straight lines.

Focussing our attention on equation (3.14), we mention some of its obvious solutions:

- (i) $V = V(x)$ or $V = V(y)$;
- (ii) $V = V(k_1 x + k_2 y)$, k_1 and k_2 constants;
- (iii) $V = x^2 + y^2$.

In case (i), equation (3.12) can be satisfied only for the trivial case, $V = \text{const}$, as follows from Remark 3.1. For the class of potentials in case (ii), we obtain from (3.12) the family of straight lines given

by $\gamma = -\frac{k_1}{k_2} = \text{const}$. In case (iii), the family of straight lines has a homogeneous of order 0 slope function $\gamma = -\frac{x}{y}$. The special solution $V = x^2 + y^2$ allows us to state that all central potentials $V = V(r)$ are solutions of (3.14).

Other classes of potentials, some expressed in polar coordinates, are given in Bozis and Anisiu (2001).

4. Equations of the direct problem

Bozis' equation represents a relation between the function γ and the potential V . It can be used to face the direct problem of Dynamics: given a potential V , find the families of orbits which can be generated. The nonlinear second order differential equation relating potentials and orbits in the form suitable for the direct problem (Bozis, 1995) is

$$\begin{aligned} \gamma^2 \gamma_{xx} - 2\gamma \gamma_{xy} + \gamma_{yy} &= \frac{\gamma \gamma_x - \gamma_y}{V_y \gamma + V_x}. \\ (-\gamma_x V_x + (2\gamma \gamma_x - 3\gamma_y) V_y + \gamma (V_{xx} - V_{yy}) + (\gamma^2 - 1) V_{xy}) . \end{aligned} \quad (4.15)$$

This is obtained by rearranging Bozis' equation (2.9).

Due to its nonlinearity in γ , it is difficult to be solved. Additional information may help in searching for solutions. The case of families produced by homogeneous potentials was considered by Bozis and Stasiades (1993), and by Bozis and Grigoriadou (1993); the problem was reduced to solving ordinary differential equations. Homogeneous families produced by inhomogeneous potentials were studied by Bozis et al (1997), as well as families of orbits with $\gamma = \gamma(x)$, corresponding to families $f(x, y) = y + h(x) = c$ (Bozis et al, 2000); in these two cases γ was found as the common root of some algebraic equations in γ , with coefficients depending on V and on its derivatives.

The additional condition satisfied by γ may be put in the terms of a first order differential equation. Indeed, if f is homogeneous of degree m , then γ is homogeneous of degree 0. This happens if and only if $x\gamma_x + y\gamma_y = 0$. For the family $f(x, y) = y + h(x)$ the corresponding γ is given by $\gamma = \frac{1}{h'(x)}$ and satisfies the equation $\gamma_y = 0$.

More generally, we can suppose that we have additional information on the family of curves (2.1) given as a linear first order differential equation which is satisfied by γ , $a(x, y)\gamma_x + b(x, y)\gamma_y = 0$. In this case, as in the special cases mentioned above, if the potential satisfies a differential condition, the family γ can be obtained as a common solution of two polynomial equations of degree at most seven, respectively twelve (Bozis et al, 2002). The coefficients of the polynomials in γ are expressions containing the derivatives of V up to the forth order, and can be calculated using symbolic algebra programs.

Example 4.1 For the Hénon-Heiles potential $V(x, y) = \frac{1}{2}x^2 + 8y^2 + x^2y + \frac{16}{3}y^3$ and $a(x, y) = x$, $b(x, y) = y$, the solution $\gamma = -\frac{x}{4y}$ corresponding to the family $yx^{-4} = c$ was found by Bozis and al (1997) as an example of a homogeneous family traced under the action of an inhomogeneous potential. The energy on the family is given by $E = -\frac{1}{24c}$ and the allowed region is $(x^2 + 8y^2 + 12y)y \leq 0$.

5. Programmed motion

It was proved in section 2 that during the motion of a material point of unit mass along an orbit of the family (2.1), the inequality

$$B(x, y) \geq 0 \quad (5.16)$$

must be observed, with

$$B = E(f(x, y)) - V(x, y). \quad (5.17)$$

This means that the motion is allowed along those members of the family (2.1) which are lying only inside some regions of the xy plane. The function $B(x, y)$ is the kinetic energy of the material point of unit mass, as it moves on any of the orbits in the presence of the potential $V(x, y)$.

According to Galliulin (1984), dynamical systems with programmed motion “are solved in such a way that the process occurring in these systems satisfies some preset requirements”. The requirement that the motion takes place in the region (5.16) was considered at first by Bozis (1996). This type of programmed motion is the following: Given a preassigned region in the plane defined by (5.16), find a potential V which produces as trajectories of (2.2) the curves in the family (2.1).

Anisiu and Bozis (2000) studied this problem completely for a simpler family of functions, given by

$$f(x, y) = y - h(x) = c, \quad (5.18)$$

where h is a nonlinear ($h''(x) \neq 0$) function of x . A simpler function b , which is nonnegative if and only if B is nonnegative, was considered. Under certain conditions on b , the function h (and, consequently, the family f) as well as the energy dependence function $E(f)$ and the potential $V(x, y)$ were determined. Let us denote

$$b = B/(1 + \gamma^2). \quad (5.19)$$

The function b satisfies a second order partial differential equation for the general family of functions (2.1). This was derived at first by Bozis (1995) starting from the case of nonconservative forces. Here we shall obtain it using Szebehely’s equation.

Theorem 5.1 Denoting $\kappa = \frac{1}{\gamma} - \gamma$, $m = 2\gamma_y + \frac{\gamma_x}{\gamma}$, $n = 2\gamma_x - \frac{3\gamma_y}{\gamma}$, $p = \frac{2\Gamma_y}{\gamma}$, the function b satisfies the linear second order partial differential equation (Bozis, 1995)

$$-b_{xx} + \kappa b_{xy} + b_{yy} = mb_x + nb_y + pb. \quad (5.20)$$

Proof. In view of (5.17) and (5.19), Szebehely's equation (2.7) becomes

$$V_x + \gamma V_y = -2\Gamma b, \quad (5.21)$$

and $E = V + (1 + \gamma^2)b$. But $\gamma = \frac{E_y}{E_x}$, so we obtain $\gamma = \frac{V_y + ((1 + \gamma^2)b)_y}{V_x + ((1 + \gamma^2)b)_x}$, which can be written as

$$\gamma V_x - V_y = (1 + \gamma^2)(b_y - \gamma b_x) - 2\gamma\Gamma b. \quad (5.22)$$

Solving the system of two equations (5.21) and (5.22) we get $V_x = \gamma(b_y - \gamma b_x) - 2\Gamma b$, $V_y = \gamma b_x - b_y$. Writing the compatibility condition $(V_x)_y = (V_y)_x$ we obtain (5.20). ■

We can rearrange (5.20) as a nonlinear equation of order two in γ

$$\begin{aligned} 2b\gamma\gamma_{xy} - 2b\gamma_{yy} &= -2b\gamma_x\gamma_y - (2b_y\gamma + b_x)\gamma_x \\ &- (2b_x\gamma - 3b_y)\gamma_y - b_{xy}\gamma^2 + (b_{yy} - b_{xx})\gamma + b_{xy}, \end{aligned} \quad (5.23)$$

which can be used in the case when the function b is known.

Considering the kinetic energy $B = 1$ (or, equivalently, $B = \text{const}$) for all the curves of the family (2.1), equation (5.20) will become

$$(1 + \gamma^2)(\gamma^2\gamma_{xx} - 2\gamma\gamma_{xy} + \gamma_{yy}) + 2\gamma(1 - \gamma^2)\gamma_x^2 - 4\gamma\gamma_y^2 + 2(3\gamma^2 - 1)\gamma_x\gamma_y = 0 \quad (5.24)$$

which is the equation giving the totality of isotach orbits (Bozis, 1986).

For the special case of the family (5.18), the function z in Anisiu and Bozis (2000) is equal to $-\gamma$, and equation (5.20) becomes

$$(b_x - 2b_y z) z' = b_{xy} z^2 + (b_{yy} - b_{xx}) z - b_{xy},$$

which was derived there directly. The method presented in the cited paper gives the family of orbits, the energy and the potential, when a suitable function b defines the allowed region.

Example 5.1 For $x > 0$ and the function $b = b_2 y^2 + b_1 y + b_0$, with $b_2 = -3/x$, $b_1 = 3(x^4 + 1)/x^2$, $b_0 = -(3x^4(x^4/5 + 1) + 1)/x^3$, the family of orbits of the form (5.18) $y - \frac{1}{x} = c$ is described inside the region $b(x, y) \geq 0$. In this case, $E = c^3$ and the potential is $V(x, y) = y^3 + 3x^3y^2 - 3x^2(x^4 + 2)y + \frac{1}{5}x(3x^8 + 18x^4 + 20)$.

6. Final remarks

PDEs appear in connection with the inverse problem of dynamics; (2.7) is quasilinear and (2.9) is linear in the potential function V . The direct problem gives rise to the nonlinear equation (4.15). Other PDEs are produced by related problems: the study of potentials creating families of straight lines leads to equation (3.14), the programming of the

motion in certain regions of the plane to (5.20), (5.23), and the study of families traced with constant kinetic energy to (5.24).

Acknowledgement. This research was partially supported by the Ministry of Education and Research, by grant 343-CNCSIS 33444/2002.

References

- [1] Anisiu, M.-C. : 1998, *Nonlinear Analysis Methods with Application in Celestial Mechanics*, University Press, Cluj-Napoca (in Romanian)
- [2] Anisiu, M.-C. and Bozis, G. : 2000, Programmed motion for a class of families of planar orbits, *Inverse Problems* **16**, 19-32
- [3] Bozis, G. : 1983, Inverse problem with two-parametric families of planar orbits, *Celest. Mech.* **31**, 129-143
- [4] Bozis, G. : 1984, Szebehely's inverse problem for finite symmetrical material concentrations, *Astronom. Astrophys.* **134**, 360-364
- [5] Bozis, G. : 1986, Adelphic potentials, *Astron. Astrophys.* **160**, 107-110
- [6] Bozis, G. and Grigoriadou, S. : 1993, Families of planar orbits generated by homogeneous potentials, *Celest. Mech.* **57**(3), 461-472
- [7] Bozis, G. and Stefiades, Ap.: 1993, Geometrically similar orbits in homogeneous potentials, *Inverse Problems* **9**(2), 233-240
- [8] Bozis, G. and Ichtiaroglou, S.: 1994, Boundary curves for families of planar orbits, *Celest. Mech.* **58**, 371-385
- [9] Bozis, G. : 1995, The inverse problem of dynamics: basic facts, *Inverse Problems* **11**, 687-708
- [10] Bozis, G.: 1996, Two-dimensional programmed motion, Proceedings of the 2nd Hellenic Astronomical Conference, Thessaloniki, June 29-July 1, 1995 (eds. M. E. Contadakis et al), pp. 587-590
- [11] Bozis, G., Anisiu, M.-C. and Blaga, C. : 1997, Inhomogeneous potentials producing homogeneous orbits, *Astron. Nachr.* **318**, 313-318
- [12] Bozis, G., Anisiu, M.-C. and Blaga, C. : 2000, A solvable version of the direct problem of dynamics, *Rom. Astronom. J.* **10**(1), 59-70
- [13] Bozis, G. and Anisiu, M.-C. : 2001, Families of straight lines in planar potentials, *Rom. Astronom. J.* **11**(1), 27-43
- [14] Bozis, G., Anisiu, M.-C. and Blaga, C. : 2002, Special families of orbits in the direct problem of dynamics, preprint
- [15] Broucke, R. and Lass, H. : 1977, On Szebehely's equation for the potential of a prescribed family of orbits, *Celest. Mech.* **16**, 215-225
- [16] Galiullin, A. S. : 1984, *Inverse Problem of Dynamics*, p. 91, Mir Publishers, Moscow
- [17] Grigoriadou, S., Bozis, G. and Elmabsout, B.: 1999, Solvable cases of Szebehely's equation, *Celest. Mech.* **74**, 211-221
- [18] Newton, I. : 1687, *Philosophiae Naturalis Principia Mathematica*, London
- [19] Szebehely, V. : 1974, On the determination of the potential by satellite observations, in E. Proverbio (ed.) *Proc. Intern. Meeting on Earth's Rotations by Satellite Observations*, The University of Cagliari, Bologna, Italy, 31-35.

EXISTENCE AND ASYMPTOTIC BEHAVIOR FOR SOME DIFFERENCE EQUATIONS ASSOCIATED WITH ACCRETIVE OPERATORS

N. C. Apreutesei

Department of Mathematics

Technical University of Iasi

11, Bd. Copou, 6600, Iasi, Romania

ndumitri@tuiasi.ro

Abstract We establish the existence, uniqueness and asymptotic behavior of the solution to a class of difference equations in a real Banach space, namely (1.1) below. The operator A which governs the problem is m-accretive. This equation is of interest because it is the discrete analog of a class of evolution equations studied by many mathematicians.

Keywords: Accretive operator, m-accretive operator, strongly accretive operator, strongly monotone duality mapping.

1. Introduction

We are concerned with the difference equation

$$\begin{cases} u_{i+1} - (1 + \theta_i) u_i + \theta_i u_{i-1} \in c_i A u_i, & i \geq 1 \\ u_0 = a, & \sup_{i \geq 1} \|u_i\| < \infty, \end{cases} \quad (1.1)$$

where A is a nonlinear m-accretive (possibly multivalued) operator in a real Banach space $(X, \|\cdot\|)$, $a \in X$ is a given element, $\theta_i \geq 1$, $c_i > 0$, $(\forall i \geq 1)$ are two given sequences of real numbers. The existence of the solution to (1.1) and some convergence properties of the solution are investigated in this paper.

G. Morosanu [12] established the existence and uniqueness of the solution for the boundary value problem

$$\begin{cases} u_{i+1} - 2u_i + u_{i-1} \in c_i A u_i, & i \geq 1 \\ u_0 = a, & \sup_{i \geq 0} \|u_i\| < \infty, \end{cases} \quad (1.2)$$

which corresponds to the case $\theta_i \equiv 1$ in (1.1). E. Mitidieri and G. Morosanu [11] analyzed the asymptotic behavior of the solution of problem (1.2). Problem (1.2) was considered in a Hilbert space, where A is a maximal monotone operator. It is the discrete analog of the boundary value problem

$$\begin{cases} u'' \in Au, & t \in (0, \infty) \\ u(0) = a, \sup_{t \geq 0} \|u(t)\| < \infty, \end{cases} \quad (1.3)$$

which was studied by V. Barbu [7], [8]. The equation of problem (1.3) with the condition $u'(0) \in \partial j(u(0) - a)$ instead of $u(0) = a$ (where function $j : H \rightarrow (-\infty, +\infty]$ is convex, lower-semicontinuous and proper and ∂j is its subdifferential), was investigated by H. Brézis [10].

The existence and asymptotic behavior for (1.2) in the case when A is an m-accretive operator in a Banach space, were proved by E. Poffald and S. Reich [14], [15] and by S. Reich and I. Shafrir [16].

A generalization of equation (1.3) is

$$pu'' + ru' \in Au + f, \quad t \in (0, T) \quad (T \leq \infty). \quad (1.4)$$

In both cases $T < \infty$ and $T = \infty$, functions p and r are in $W^{1,\infty}(0, T)$. Papers concerned with this equation with different boundary conditions are due to L. Véron [17], N. Pavel [13], A. Aftabizadeh and N. Pavel [1], [2], N. Apreutesei [3], [4], [5].

A discretization of (1.4) with $f \equiv 0$ is

$$p_i(u_{i+1} - 2u_i + u_{i-1}) + r_i(u_{i+1} - u_i) \in k_i Au_i, \quad i \geq 1,$$

with $p_i \geq c > 0$, $k_i > 0$, $(\forall) i \geq 1$. Denoting by $\theta_i = \frac{p_i}{p_i + r_i}$, $c_i = \frac{k_i}{p_i + r_i}$, we find the equation of (1.1). Suppose that $(\theta_i)_{i \geq 1}$ is nonincreasing, $\theta_i \geq 1$, for all i . In Hilbert spaces, this equation was studied by N. Apreutesei [6].

Recall some notions we need in the following sections.

Let X be a real Banach space with the norm $\|\cdot\|$, X^* its dual space and (\cdot, \cdot) the pairing between X and X^* . Denote by $J : X \rightarrow X^*$,

$$J(x) = \{x^* \in X^*, (x, x^*) = \|x\|^2 = \|x^*\|^2\} \quad (1.5)$$

the duality mapping of X . It is obvious that J is monotone. J is single-valued if and only if X is smooth. In this case we say that J is strongly monotone if there is a positive constant M such that

$$(x - y, Jx - Jy) \geq M\|x - y\|^2, \quad (\forall) x, y \in X. \quad (1.6)$$

A subset A of $X \times X$ with the domain $D(A)$ and the range $R(A)$ is called *accretive* if for any $y_i \in Ax_i$, $i = 1, 2$, there exists $j \in J(x_1 - x_2)$ such that

$$(y_1 - y_2, j) \geq 0. \quad (1.7)$$

The accretive operator $A \subset X \times X$ is *m-accretive* if $R(I + A) = X$, where I is the identity operator of X . It follows that $R(I + \lambda A) = X$, $(\forall) \lambda > 0$. The operator $A \subset X \times X$ is said to be *strongly accretive* if there is a constant $\omega > 0$ with the property: $(\forall)y_i \in Ax_i$, $i=1, 2$, $(\exists)j \in J(x_1 - x_2)$, such that

$$(y_1 - y_2, j) \geq \omega \|x_1 - x_2\|^2. \quad (1.8)$$

It is known that if $A \subset X \times X$ is m-accretive, then A is closed. If in addition X^* is uniformly convex, then A is demiclosed (strongly-weakly closed in $X \times X$).

Section 2 is dedicated to the existence of the solution to problem (1.1). First we establish the existence for an auxiliary finite difference equation and then we use this result to the study of the problem (1.1). If X has a strongly monotone duality mapping, then (1.1) has a unique solution, for any $a \in X$. Section 3 is concerned with the asymptotic behavior of the solution. First we give a weak convergence theorem. The Hilbert spaces case is separately studied. Finally we establish a strong convergence result under the hypothesis that A is m-accretive and strongly accretive.

2. Existence results

In this section we shall study the existence and uniqueness of the solution for the difference inclusion (1.1).

We begin with an existence result for the auxiliary problem

$$\begin{cases} u_{i+1} - (1 + \theta_i) u_i + \theta_i u_{i-1} \in c_i A u_i + f_i, & 1 \leq i \leq N \\ u_0 = a, & u_{N+1} = b, \end{cases} \quad (2.1)$$

where N is a positive integer, $c_i > 0$, $\theta_i \geq 1$ and $f_i \in X$, $1 \leq i \leq N$. We work in the product space $X^N = X \times \dots \times X$ provided with the norm

$$|u| = \left(\sum_{i=1}^N \|u_i\|^2 \right)^{1/2}, \quad (\forall) u = (u_1, \dots, u_N) \in X^N. \quad (2.2)$$

If $u = (u_1, \dots, u_N) \in X^N$ and $u^* = (u_1^*, \dots, u_N^*) \in (X^*)^N$, then we denote

$$(u, u^*)_N = \sum_{i=1}^N (u_i, u_i^*). \quad (2.3)$$

Theorem 2.1. *If X is a Banach space, $A \subset X \times X$ is an m -accretive operator, $(\theta_i)_{1 \leq i \leq N}$ a nonincreasing sequence, $\theta_i > 0$, $c_i > 0$, $1 \leq i \leq N$, a, b are in X and $(f_i)_{1 \leq i \leq N} \in X^N$, then (2.1) has a unique solution in X^N .*

Proof. Denote by $\mathcal{A} \subset X^N \times X^N$ the operator

$$\mathcal{A}u = \{(c_1 v_1, \dots, c_N v_N), v_i \in Au_i, 1 \leq i \leq N\} + (\theta_1 a, 0, \dots, 0, b),$$

where $u = (u_1, \dots, u_N) \in D(A)^N$ and by $B : X^N \rightarrow X^N$ the operator

$$Bu = ((1 + \theta_1)u_1 - u_2, -\theta_2 u_1 + (1 + \theta_2)u_2 - u_3, \dots, -\theta_{N-1} u_{N-2} + (1 + \theta_{N-1})u_{N-1} - u_N, -\theta_N u_{N-1} + (1 + \theta_N)u_N).$$

The operator \mathcal{A} is m -accretive in X^N and B is continuous, everywhere defined and strongly accretive. Indeed, since B is linear, to show the strong accretivity of B it is sufficient to prove that $(\exists) \alpha > 0$, such that

$$(Bu, u^*)_N \geq \alpha |u|^2, \quad (\forall) u \in X^N, \quad (\forall) u^* \in Ju. \quad (2.4)$$

But

$$(Bu, u^*)_N = \sum_{i=1}^N (1 + \theta_i) (u_i, u_i^*) - \sum_{i=1}^{N-1} [(u_{i+1}, u_i^*) + \theta_{i+1} (u_i, u_{i+1}^*)],$$

where $u_i^* \in Ju_i$, $u_{i+1}^* \in Ju_{i+1}$. Since $(\theta_i)_{1 \leq i \leq N}$ is nonincreasing, by

$$2(u_{i+1}, u_i^*) \leq \gamma_i \|u_{i+1}\|^2 + (1/\gamma_i) \|u_i\|^2, \quad 2(u_i, u_{i+1}^*) \leq \gamma_i \|u_{i+1}\|^2 + (1/\gamma_i) \|u_i\|^2,$$

for all $\gamma_i > 0$, we find

$$(Bu, u^*)_N \geq \sum_{i=1}^N \alpha_i \|u_i\|^2, \quad (2.5)$$

with

$$\begin{cases} \alpha_1 = 1 + \theta_1 - \frac{1}{2\gamma_1} (1 + \theta_2), \\ \alpha_i = 1 + \theta_i - \frac{1}{2\gamma_i} (1 + \theta_{i+1}) - \frac{\gamma_{i-1}}{2} (1 + \theta_i), \quad 2 \leq i \leq N-1, \\ \alpha_N = (1 + \theta_N) \left(1 - \frac{\gamma_{N-1}}{2}\right). \end{cases}$$

Taking, for example $\gamma_i = i(i+2)/(i+1)^2$, we get $\alpha_i > 0$, $1 \leq i \leq N$. Denoting $\alpha = \min \{\alpha_i, 1 \leq i \leq N\} > 0$, from (2.5) one deduces

$$(Bu, u^*)_N \geq \alpha |u|^2, \quad (2.6)$$

i.e. B is strongly accretive. This implies that $\mathcal{A}+B$ is m-accretive and coercive, and consequently surjective: $R(\mathcal{A}+B) = X^N$. This means that problem (2.1) has a solution. It is easy to show the uniqueness of the solution.

Now we are going to study the difference equation (1.1), supposing that X has a strongly monotone duality mapping J . For $\theta_i \equiv 1$, this problem was studied by G. Morosanu [12] in Hilbert spaces and by E. Poffald and S. Reich [14] in Banach spaces. Recall that J is strongly monotone if and only if X is uniformly convex with a modulus of convexity of power type 2 ([14]).

We state the following existence and uniqueness result.

Theorem 2.2. *Let X be a Banach space with a strongly monotone duality mapping J and $A \subset X \times X$ an m-accretive operator, with $0 \in R(A)$. Let $(c_i)_{i \geq 1}$, $(\theta_i)_{i \geq 1}$ be two sequences, $c_i > 0$, $\theta_i \geq 1$, $(\forall) i \geq 1$, θ_i nonincreasing. Then, for every $a \in X$, problem (1.1) has a unique solution $(u_i)_{i \geq 1}$, with $u_i \in D(A)$ for all $i \geq 1$.*

Proof. By Theorem 2.1, the sequence of approximating problems

$$\begin{cases} u_{i+1}^N - (1 + \theta_i) u_i^N + \theta_i u_{i-1}^N \in c_i A u_i^N, & 1 \leq i \leq N \\ u_0^N = u_{N+1}^N = a, \end{cases} \quad (2.7)$$

has a unique solution $(u_i^N)_{1 \leq i \leq N} \in X^N$.

For a given $w \in A^{-1}(0)$, we set $w_i^N = u_i^N - w$, $0 \leq i \leq N+1$. By the accretivity of A , for every $i \in \{1, \dots, N\}$, there is a $j_i \in J(w_i^N)$, such that

$$(w_{i+1}^N - (1 + \theta_i) w_i^N + \theta_i w_{i-1}^N, j_i) \geq 0.$$

This implies

$$\|w_i^N\| \leq \frac{1}{1 + \theta_i} \|w_{i+1}^N\| + \frac{\theta_i}{1 + \theta_i} \|w_{i-1}^N\|, \quad 1 \leq i \leq N, \quad (2.8)$$

hence

$$\|w_i^N\| = \|u_i^N - w\| \leq \max(\|w_0^N\|, \|w_{N+1}^N\|) = \|a - w\|, \quad 1 \leq i \leq N$$

and thus

$$\|u_i^N\| \leq \|w\| + \|a - w\|, \quad 1 \leq i \leq N. \quad (2.9)$$

We prove now the convergence of u_i^N as $N \rightarrow \infty$ (uniformly for i belonging to every finite set of natural numbers) to an element u_i which verifies (1.1).

Let $N_0 < N_1 < N_2$ be positive integers and $v_i = u_i^{N_2} - u_i^{N_1}$, $0 \leq i \leq N_1 + 1$. Since A is accretive, for each i there is a $l_i \in Jv_i$ such that

$$(v_{i+1} - (1 + \theta_i) v_i + \theta_i v_{i-1}, l_i) \geq 0.$$

Hence

$$(v_{i+1} - v_i, l_i) - \theta_i(v_i - v_{i-1}, l_{i-1}) \geq \theta_i(v_i - v_{i-1}, l_i - l_{i-1}) \quad (2.10)$$

and, since J is strongly monotone (say of constant M), we get

$$M\theta_i\|v_i - v_{i-1}\|^2 \leq (v_{i+1} - v_i, l_i) - \theta_i(v_i - v_{i-1}, l_{i-1}), \quad (2.11)$$

for $1 \leq i \leq N_1$. One multiplies (2.11) by $\theta_k \dots \theta_{i+1}$ and one sums from $i = 1$ to $i = k$, with $k \in \{1, \dots, N_1\}$, to find

$$M \sum_{i=1}^k \theta_k \dots \theta_{i+1} \theta_i \|v_i - v_{i-1}\|^2 \leq (v_{k+1} - v_k, l_k) - \theta_k \dots \theta_1 (v_1 - v_0, l_0).$$

Since $l_0 = 0$ and $\theta_i \geq 1$, $i \geq 1$, it follows

$$M \sum_{i=1}^k \|v_i - v_{i-1}\|^2 \leq \frac{1}{2} (\|v_{k+1}\|^2 - \|v_k\|^2), \quad 1 \leq k \leq N_1. \quad (2.12)$$

Since $\|v_k\| = \sum_{i=1}^k (\|v_i\| - \|v_{i-1}\|) \leq \sum_{i=1}^k \|v_i - v_{i-1}\|$, (2.12) implies

$$\|v_k\|^2 \leq \frac{k}{2M} (\|v_{k+1}\|^2 - \|v_k\|^2). \quad (2.13)$$

We sum from $k = N_0$ to $k = N_1$, to get with the aid of (2.9)

$$\sum_{k=N_0}^{N_1} \frac{1}{k} \|v_k\|^2 \leq \frac{1}{2M} (\|v_{N_1+1}\|^2 - \|v_{N_0}\|^2) \leq C. \quad (2.14)$$

By (2.13) we have $\|v_k\| \leq \|v_{k+1}\|$, for all $1 \leq k \leq N_1$. Hence for every $i \leq N_0$,

$$\|v_i\|^2 \left(\sum_{k=N_0}^{N_1} \frac{1}{k} \right) \leq \sum_{k=N_0}^{N_1} \frac{1}{k} \|v_k\|^2 \leq C, \quad (2.15)$$

so

$$\|u_i^{N_2} - u_i^{N_1}\|^2 \leq C / \left(\sum_{k=N_0}^{N_1} \frac{1}{k} \right), \quad 1 \leq i \leq N_0. \quad (2.16)$$

Therefore, there exists the limit $u_i = \lim_{N \rightarrow \infty} u_i^N$, for all i belonging to every finite set of natural numbers. Since A is m-accretive in X , we may pass to the limit in (2.7) to deduce that $(u_i)_{i \geq 1}$ is a solution of problem (1.1). The uniqueness follows easily in the same manner.

3. Asymptotic behavior

In this section, we give some weak convergence and strong convergence results for the solution of the problem (1.1). Let's begin with an auxiliary result.

Proposition 3.1. *Under the hypotheses of Theorem 2.2, if $(u_i)_{i \geq 1}$ is the unique solution of problem (1.1), then $(\forall) w \in A^{-1}(0)$, the sequence $(|u_i - w|)_{i \geq 1}$ is nonincreasing. Moreover, there exists the limit*

$$h(w) = \lim_{i \rightarrow \infty} |u_i - w|. \quad (3.1)$$

Proof. We pass to the limit as $N \rightarrow \infty$ in (2.8), where $w_i^N = u_i^N - w$, $1 \leq i \leq N$. Since the limit $u_i = \lim_{N \rightarrow \infty} u_i^N$ exists uniformly for i belonging to every finite set of natural numbers, we get

$$|u_i - w| \leq \frac{1}{1 + \theta_i} |u_{i+1} - w| + \frac{\theta_i}{1 + \theta_i} |u_{i-1} - w|, \quad i \geq 1. \quad (3.2)$$

Since $(u_i)_{i \geq 1}$ is bounded, this inequality shows that $(|u_i - w|)_{i \geq 1}$ is nonincreasing. The monotonicity and the boundedness of $|u_i - w|$ lead us to the conclusion that the limit (3.1) exists.

Definition 3.1. The multivalued operator $A \subset X \times X$ is said to be *injectiv* if $Ax_1 \cap Ax_2 \neq \emptyset$ implies $x_1 = x_2$.

We state now a weak convergence theorem.

Theorem 3.1. *Let X be a Banach space with a strongly monotone duality mapping J and let its dual X^* be uniformly convex. Suppose $a \in X$ is a given element and $A \subset X \times X$ is injectiv and m -accretive, $0 \in R(A)$. Let $(c_i)_{i \geq 1}$, $(\theta_i)_{i \geq 1}$ be like in Theorem 2.2, $\theta_1 \dots \theta_i / c_i \leq K$, $(\forall) i \geq 1$ (K is a positive constant). Then, the solution $(u_i)_{i \geq 1}$ of problem (1.1) converges weakly as $i \rightarrow \infty$ to a zero of A .*

Proof. Denoting by $(\alpha_i)_{i \geq 1}$ the sequence $\alpha_0 = 1$, $\alpha_i = 1/\theta_1 \dots \theta_i$, $(\forall) i \geq 1$, observe that $\alpha_{i-1}/\alpha_i = \theta_i$, so

$$u_{i+1} - (1 + \theta_i) u_i + \theta_i u_{i-1} = \frac{1}{\alpha_i} [\alpha_i (u_{i+1} - u_i) - \alpha_{i-1} (u_i - u_{i-1})].$$

Thus problem (1.1) can be written as

$$\begin{cases} \frac{1}{\alpha_i c_i} (\varphi_{i+1} - \varphi_i) \in Au_i, & i \geq 1 \\ u_0 = a, & \sup_{i \geq 1} \|u_i\| < \infty, \end{cases} \quad (3.3)$$

where $\varphi_i = \alpha_{i-1} (u_i - u_{i-1}) = \alpha_{i-1} (w_i - w_{i-1})$. Here $w_i = u_i - w$, with a fixed $w \in A^{-1}(0)$. Since A is accretive, for every i , there is $j_i \in Jw_i$ such as

$$(w_{i+1} - (1 + \theta_i) w_i + \theta_i w_{i-1}, j_i) \geq 0.$$

This is equivalent with

$$\begin{aligned} \frac{1}{\alpha_i} [(\varphi_{i+1}, j_i) - (\varphi_i, j_{i-1})] &\geq \frac{1}{\alpha_i} (\varphi_i, j_i - j_{i-1}) = \\ &= \frac{\alpha_{i-1}}{\alpha_i} (u_i - u_{i-1}, j_i - j_{i-1}) \end{aligned} \quad (3.4)$$

and since J is strongly accretive (of constant M), for every positive integer N , we get

$$\begin{aligned} M \sum_{i=1}^N \alpha_{i-1} \|u_i - u_{i-1}\|^2 &\leq \sum_{i=1}^N [(\varphi_{i+1}, j_i) - (\varphi_i, j_{i-1})] = \\ &= \frac{\alpha_N}{2} (\|w_{N+1}\|^2 - \|w_N\|^2) + \frac{1}{2} \|w_1\|^2 + \frac{3}{2} \|w_0\|^2. \end{aligned} \quad (3.5)$$

Using the estimation for w_i in the proof of Theorem 2.2, one obtains

$$M \sum_{i=1}^N \alpha_{i-1} \|u_i - u_{i-1}\|^2 \leq \frac{5}{2} \|a - w\|^2. \quad (3.6)$$

Passing to the limit as $N \rightarrow \infty$, one finds that $\alpha_{i-1} \|u_i - u_{i-1}\|^2 \rightarrow 0$. Since $\alpha_{i-1} \leq 1$, we have $\alpha_{i-1} (u_i - u_{i-1}) \rightarrow 0$ strongly in X .

Let u be a weak limit of a weakly convergent subsequence (u_{i_n}) of $(u_i)_{i \geq 1}$. By the assumption $\theta_1 \dots \theta_i / c_i \leq K$, $(\forall) i \geq 1$, since A is demiclosed, we may pass to the limit in (3.3) (as $i_n \rightarrow \infty$) and obtain $0 \in Au$.

If u , \tilde{u} are two such weak limits of some weakly convergent subsequences of $(u_i)_{i \geq 1}$, then from the injectivity of A , we deduce that $u = \tilde{u}$. So there is an element $u \in A^{-1}(0)$ such that $u_i \rightharpoonup u$, $i \rightarrow \infty$.

In Hilbert spaces, we have the following result, which for $\theta_i \equiv 1$ was proved by G. Morosanu [12].

Theorem 3.2. *Let H be a real Hilbert space, $a \in H$ be given and $A : D(A) \subset H \rightarrow H$ be a maximal monotone operator in H , with $0 \in R(A)$. If $(c_i)_{i \geq 1}$, $(\theta_i)_{i \geq 1}$ satisfy the conditions of Theorem 3.1 and, in addition, there is a constant $c > 0$ such that $c_i \geq c > 0$, then the solution u_i of problem (1.1) is weakly convergent to a zero of A .*

We give now a strongly convergence result. The proof is similar to the Hilbert spaces case (see N. Apreutesei [6]).

Theorem 3.3. *Let X be a Banach space with a strongly monotone duality mapping J , $a \in X$ be a given element and $A \subset X \times X$ an m -accretive, strongly accretive and univoque operator with $0 \in R(A)$. Let $(c_i)_{i \geq 1}$, $(\theta_i)_{i \geq 1}$ be two sequences such that $(\theta_i)_{i \geq 1}$ is nonincreasing, $c_i > 0$, $\theta_i \geq 1$, $(\forall) i \geq 1$ and $\sum_{i=1}^{\infty} c_i / \theta_i = \infty$. Then, the solution $(u_i)_{i \geq 1}$ of problem (1.1) is strongly convergent as $i \rightarrow \infty$ to the only element of $A^{-1}0$.*

References

- [1] A.R. Aftabizadeh and N.H. Pavel, Boundary value problems for second order differential equations and a convex problem of Bolza, *Diff. Integral Eqns.* **2**(1989), 495-509.
- [2] A.R. Aftabizadeh and N.H. Pavel, Nonlinear boundary value problems for some ordinary and partial differential equations associated with monotone operators, *J.Math.Anal.Appl.* **156**(1991), 535-557.
- [3] N.C. Apreutesei, A boundary value problem for second order differential equations in Hilbert spaces, *Nonlinear Analysis, TMA*, **24**(1995), 1235-1246.
- [4] N.C. Apreutesei, Some second order evolution equations governed by maximal monotone operators, *Anal. Univ. Craiova*, **24**(1997), 45-61.
- [5] N.C. Apreutesei, Second order differential equations on half-line associated with monotone operators, *J.Math.Anal.Appl.* **223**(1998), 472-493.
- [6] N.C. Apreutesei, Existence and asymptotic behavior for a class of second order difference equations, to appear in *J. Difference Eq. Appl.*
- [7] V. Barbu, Sur un probleme aux limites pour une classe d'équations différentielles nonlinéaires abstraites du deuxième ordre en t , *C.R.Acad.Sci. Paris* **274**(1972), 459-462.
- [8] V. Barbu, A class of boundary problems for second order abstract differential equations, *J.Fac.Sci.Univ.Tokyo,Sect 1*, **19**(1972), 295-319.
- [9] V. Barbu, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Leyden, 1976.
- [10] H. Brézis, Equations d'évolution du second ordre associees à des opérateurs monotones, *Israel J. Math.* **12**(1972), 51-60.
- [11] E. Mitidieri and G. Moroşanu, Asymptotic behaviour of the solutions of second order difference equations associated to monotone operators, *Numerical Funct. Anal. Optim.* **8**(1986-1987), 419-434.
- [12] G. Moroşanu, Second order difference equations of monotone type, *Numerical Funct. Anal. Optim.* **1**(1979), 441-450.
- [13] N. Pavel, Nonlinear boundary value problems for second order differential equations, *J. Math. Anal. Appl.* **50**(1975) 373-383.

- [14] E. Poffald and S. Reich, An incomplete Cauchy problem, *J. Math. Anal. Appl.*, **113**(1986), 514-543.
- [15] E. Poffald and S. Reich, A difference inclusion, in "Nonlinear Semigroups, Partial Differential Equations and Attractors", Lecture Notes in Mathematics, vol. **1394**, Springer, Berlin, 1989, 122-130.
- [16] S. Reich and I. Shafrir, An existence theorem for a difference inclusion in general Banach spaces, *J. Math. Anal. Appl.*, **160** (1991), 406-412.
- [17] L. Véron, Problèmes d'évolution du second ordre associés à des opérateurs monotones, *C.R. Acad.Sci.Paris* **278**(1974), 1099-1101.

CONVERGENCE RATE OF A MULTIPLICATIVE SCHWARZ METHOD FOR STRONGLY NONLINEAR VARIATIONAL INEQUALITIES

L. Badea

*Institute of Mathematics,
Romanian Academy of Sciences,
P.O. Box 1-764,
RO-70700 Bucharest, Romania*
lbadea@imar.ro

Abstract We prove the convergence and estimate the error of a general algorithm for the minimization of non-quadratic functionals over a convex set in a reflexive Banach space, provided that the convex set verifies a certain assumption. In the case of the Sobolev spaces, our algorithm is exactly a variant of the Schwarz domain decomposition method, and we prove that the introduced assumption holds if the convex set is defined by constraints on the function values almost everywhere in the domain. In the end of the paper we give some numerical examples concerning the two-obstacle problem of a nonlinear elastic membrane.

Keywords: domain decomposition methods, Schwarz method, nonlinear variational inequalities, nonlinear minimization, obstacle problems

1. Introduction

The literature on the domain decomposition methods is very large and it is motivated by an increasing need on the solution of large-scale problems since these methods provide numerical solvers which are efficient and parallelizable on multi-processor machines. However, to our knowledge, very few papers deal with the application of these methods to nonlinear problems.

The main goal of this paper is to give an error estimate for a Schwarz domain decomposition method applied to the minimization of the non quadratic functionals over a convex set which is not supposed to be decomposed as a sum of subconvex sets. The convergence of a domain de-

composition algorithm solving variational inequalities coming from the minimization of quadratic functionals over convex sets which are defined by constraints on the function values at the points of the domain is proved in [2]. In [9], it is proved that the multiplicative space decomposition method applied to the minimization without constraints of a differentiable and convex functional defined in a reflexive Banach space uniformly converges. In [3], it is proved that the method in [2] converges for the more general conditions given in [9] in the case of the quadratic functional minimization. We generalize in this paper the results in [3] and [9] to the minimization of the non quadratic functionals.

The paper is organized as follows. In Section 2, we state the multiplicative Schwarz method for nonlinear variational inequalities as a subspace correction method in a general reflexive Banach space for the minimization of non quadratic functionals, and we prove the convergence of this algorithm provided that a certain assumption holds. In Section 3, under a little stronger assumption, which essentially introduces a constant depending on the convex set and the space decomposition, we estimate the error of the algorithm. Section 4 is devoted to the convergence of the method in Sobolev spaces, proving that the introduced assumptions hold. For the Sobolev spaces, the algorithm is exactly a variant of the Schwarz method. Finally, in Section 5, we illustrate the method by numerical examples concerning the two-obstacle problem of a nonlinear elastic membrane.

2. General convergence result

Let V be a reflexive Banach space and V_1, \dots, V_m , be some closed subspaces of V . Also, we consider a non empty closed convex set $K \subset V$, and we make the following

Assumption 1. *For any $w, v \in K$ and $w_i \in V_i$ with $w + \sum_{j=1}^i w_j \in K$, $i = 1, \dots, m$, there exist $v_i \in V_i$, $i = 1, \dots, m$, satisfying*

$$w + \sum_{j=1}^{i-1} w_j + v_i \in K \text{ for } i = 1, \dots, m, \quad (1)$$

$$v - w = \sum_{i=1}^m v_i, \quad (2)$$

and the application from $V \times V_1 \times \dots \times V_m$ to $V_1 \times \dots \times V_m$

$$(v - w, w_1, \dots, w_m) \rightarrow (v_1, \dots, v_m). \quad (3)$$

is bounded, i.e. it transforms the bounded sets in some bounded sets.

This assumption looks to be complicated enough, but, as we shall see in Section 4, it holds for problems in which we use the Sobolev spaces and the convex set K is defined by constraints of the function values at the points of the domain. We consider a Gâteaux differentiable functional $F : K \rightarrow R$, which will be assumed to be coercive if K is not bounded. We assume that for any real number $M > 0$, if we write $L_M = \sup_{\|v\|, \|u\| \leq M, v, u \in K} \|v - u\|$, there exist two functions $\alpha_M, \beta_M : [0, L_M] \rightarrow \mathbf{R}^+$, such that

$$\alpha_M \text{ is continuous and strictly increasing, and } \alpha_M(0) = 0, \quad (4)$$

$$\beta_M \text{ is continuous at } 0 \text{ and } \beta_M(0) = 0, \quad (5)$$

and satisfying for any $u, v \in K$ with $\|u\|, \|v\| \leq M$,

$$\langle F'(v) - F'(u), v - u \rangle \geq \alpha_M(\|v - u\|), \quad (6)$$

$$\beta_M(\|v - u\|) \geq \|F'(v) - F'(u)\|_{V'}, \quad (7)$$

where F' is the Gâteaux derivative of F .

We know (see [5], Proposition 5.5) that if (6) holds for any $M > 0$, then the functional F is strictly convex. Also, it is evident that if (7) holds, then F is continuously differentiable. Reciprocally, we can prove in a similar way to that given in [6], Lemma 1.1, for the case of the Euclidean spaces, that if the closed unity ball is compact in the strong topology of the space Banach V , F' is continuous and F is strictly convex, then the functions $\alpha_M(\tau) = \inf_{\|v-u\|=\tau, \|v\|, \|u\| \leq M, v, u \in K} \langle F'(v) - F'(u), v - u \rangle$ and $\beta_M(\tau) = \sup_{\|v-u\|=\tau, \|v\|, \|u\| \leq M, v, u \in K} \|F'(v) - F'(u)\|_{V'}$,

exist for any $M > 0$, and they satisfy (4), (6), and (5), (7), respectively.

It is evident that if (6) and (7) hold, then for any $u, v \in K$, $\|u\|, \|v\| \leq M$, we have

$$\alpha_M(\|v - u\|) \leq \langle F'(v) - F'(u), v - u \rangle \leq \beta_M(\|v - u\|) \|v - u\|. \quad (8)$$

Following the way in [6] (Lemmas 1.1 and 1.2), we can prove that for any $u, v \in K$, $\|u\|, \|v\| \leq M$, we have

$$\begin{aligned} \langle F'(u), v - u \rangle + \lambda_M(\|v - u\|) &\leq F(v) - F(u) \leq \\ \langle F'(u), v - u \rangle + \mu_M(\|v - u\|), \end{aligned} \quad (9)$$

where

$$\lambda_M(\tau) = \int_0^\tau \alpha_M(\theta) \frac{d\theta}{\theta}, \quad \mu_M(\tau) = \int_0^\tau \beta_M(\theta) d\theta. \quad (10)$$

Now, we consider the minimization problem

$$u \in K : F(u) \leq F(v), \text{ for any } v \in K. \quad (11)$$

It is well known (see [5]) that if V is a reflexive Banach space and F is strictly convex, differentiable, and coercive if K is not bounded, then the above problem has a unique solution, and this is also the unique solution of the problem

$$u \in K : \langle F'(u), v - u \rangle \geq 0, \text{ for any } v \in K. \quad (12)$$

From (9) we see that, for a given $M > 0$ such that the solution u of (12) satisfies $\|u\| \leq M$, we have

$$\lambda_M(\|v - u\|) \leq F(v) - F(u), \text{ for any } v \in K, \|v\| \leq M. \quad (13)$$

The proposed algorithm corresponding to the subspaces V_1, \dots, V_m and the convex set K is written as follows

Algorithm 2.1. *We start the algorithm with an arbitrary $u^0 \in K$. At iteration $n + 1$, having $u^n \in K$, $n \geq 0$, we compute sequentially for $i = 1, \dots, m$, $w_i^{n+1} \in V_i$ satisfying*

$$w_i^{n+1} = \arg \min_{u^{n+\frac{i-1}{m}} + v_i \in K, v_i \in V_i} G(v_i), \quad (14)$$

with $G(v_i) = F(u^{n+\frac{i-1}{m}} + v_i)$, and then we update $u^{n+\frac{i}{m}} = u^{n+\frac{i-1}{m}} + w_i^{n+1}$.

This algorithm does not assume a decomposition of the convex set K depending on the subspaces V_i . As for problem (11), since the subspaces V_i are reflexive Banach spaces, problem (14) has a unique solution, $w_i^{n+1} \in V_i$, $u^{n+\frac{i-1}{m}} + w_i^{n+1} \in K$, and it also satisfies the variational inequality

$$\begin{aligned} & \langle F'(u^{n+\frac{i-1}{m}} + w_i^{n+1}), v_i - w_i^{n+1} \rangle \geq 0, \\ & \text{for any } v_i \in V_i, u^{n+\frac{i-1}{m}} + v_i \in K. \end{aligned} \quad (15)$$

We have the following general convergence result.

Theorem 1 *We consider that V is a reflexive Banach, V_1, \dots, V_m are some closed subspaces of V , K is a non empty closed convex subset of V , and F is Gâteaux differentiable functional on K which is assumed to be coercive if K is not bounded. If Assumption 2.1 holds, and for any $M > 0$ there exist two functions α_M and β_M satisfying (4)–(7), then, for any $i = 1, \dots, m$, $u^{n+\frac{i}{m}} \rightarrow u$, strongly in V , as $n \rightarrow \infty$, where u is the*

solution of problem (11) and $u^{n+\frac{i}{m}}$ are given by Algorithm 2.1 starting from an arbitrary given u^0 .

Proof. From (15) and (9), we have for any $n \geq 0$ and $i = 1, \dots, m$,

$$F(u^{n+\frac{i-1}{m}}) - F(u^{n+\frac{i}{m}}) \geq \lambda_M(\|w_i^{n+1}\|), \quad (16)$$

and therefore, using (11), we get for any $n \geq 0$ and $i = 1, \dots, m$, that

$$F(u) \leq F(u^{n+\frac{i}{m}}) \leq F(u^{n+\frac{i-1}{m}}) \leq F(u^0). \quad (17)$$

Taking into account the boundedness of K or the coerciveness of F , it follows that there exists a real constant $M > 0$ such that

$$\|u\| \leq M, \|u^0\| \leq M, \|u^{n+\frac{i}{m}}\| \leq M \quad \forall n \geq 0, i = 1, \dots, m. \quad (18)$$

From (16) we also get

$$F(u^n) - F(u^{n+1}) \geq \sum_{i=1}^m \lambda_M(\|w_i^{n+1}\|), \text{ for any } n \geq 0. \quad (19)$$

Consequently, from (17), the series $\sum_{n=1}^{\infty} \lambda_M(\|w_i^{n+1}\|)$ is convergent for any $i = 1, \dots, m$, and therefore

$$\|w_i^{n+1}\| \rightarrow 0, \text{ as } n \rightarrow \infty, \text{ for any } i = 1, \dots, m. \quad (20)$$

Applying Assumption 2.1 for $w = u^n$, $v = u$, and $w_i = w_i^{n+1}$, we get a decomposition u_1, \dots, u_m of $u - u^n$. From (1), we can replace v_i by u_i in (15), and we have $\langle F'(u^{n+\frac{i}{m}}) - F'(u^{n+1}), u_i - w_i^{n+1} \rangle + \langle F'(u^{n+1}), u_i - w_i^{n+1} \rangle \geq 0$. Using (2) we have $\sum_{i=1}^m \langle F'(u^{n+\frac{i}{m}}) - F'(u^{n+1}), u_i - w_i^{n+1} \rangle + \langle F'(u^{n+1}), u - u^{n+1} \rangle \geq 0$. Using this inequality, from (18), (9) and (7) we obtain

$$\begin{aligned} F(u^{n+1}) - F(u) + \lambda_M(\|u - u^{n+1}\|) &\leq \langle F'(u^{n+1}), u^{n+1} - u \rangle \\ &\leq \sum_{i=1}^m \langle F'(u^{n+\frac{i}{m}}) - F'(u^{n+1}), u_i - w_i^{n+1} \rangle \\ &= \sum_{i=1}^m \sum_{j=i+1}^m \langle F'(u^{n+\frac{j-1}{m}}) - F'(u^{n+\frac{j}{m}}), u_i - w_i^{n+1} \rangle \\ &\leq \sum_{i=1}^m \beta_M(\|w_i^{n+1}\|) \sum_{i=1}^m \|u_i - w_i^{n+1}\|. \end{aligned} \quad (21)$$

From (20) and (3) we get that the sequence $\{\sum_{i=1}^m \|u_i - w_i^{n+1}\|\}_n$ is bounded. Also, from (20) and (5) we have $\sum_{i=1}^m \beta_M(\|w_i^{n+1}\|) \rightarrow 0$ as $n \rightarrow \infty$. Consequently, $F(u^{n+1}) - F(u) \rightarrow 0$ and $\lambda_M(\|u - u^{n+1}\|) \rightarrow 0$ as $n \rightarrow \infty$. Now, from (4) and (10) it is clear that $u^n \rightarrow u$ as $n \rightarrow \infty$. ■

3. Error estimate

The error estimate essentially stands on the convergence order of the functions $\alpha_M(\tau)$ and $\beta_M(\tau)$ to zero as $\tau \rightarrow 0$. In the following we take these functions of polynomial form

$$\alpha_M(\tau) = A_M \tau^p, \quad \beta_M(\tau) = B_M \tau^{q-1}, \quad (22)$$

where $A_M > 0$, $B_M > 0$, $p > 1$ and $q > 1$ are some real constants. We have marked here that the constants A_M and B_M depend on M , and we see from (8) that we must take $p \geq q$. Now, from (10) we get

$$\lambda(\tau) = \frac{A_M}{p} \tau^p, \quad \mu(\tau) = \frac{B_M}{q} \tau^q. \quad (23)$$

Naturally, the convergence rate will depend on the spaces V_1, \dots, V_m , and we shall consider the following form of Assumption 2.1 having condition (3) slightly modified

Assumption 3.1. *There exists a constant C_0 such that for any $w, v \in K$ and $w_i \in V_i$ with $w + \sum_{j=1}^i w_j \in K$, $i = 1, \dots, m$, there exist $v_i \in V_i$, $i = 1, \dots, m$, satisfying (1), (2) and*

$$\sum_{i=1}^m \|v_i\|^p \leq C_0^p \left(\|v - w\|^p + \sum_{i=1}^m \|w_i\|^p \right). \quad (24)$$

In the case of the minimization of quadratic functionals in [3], the above assumption has been introduced for $p = 2$. The following theorem is a generalization for nonlinear inequalities of the result in [9] concerning the convergence of the method for nonlinear equations.

Theorem 2 *On the conditions of Theorem 1 we consider the functions α_M and β_M defined in (22) and we make Assumption 3.1. If u is the solution of problem (11) and u^n , $n \geq 0$, are its approximations obtained from Algorithm 2.1, then we have the following error estimations:*

(i) if $p = q$ we have

$$\begin{aligned} F(u^n) - F(u) &\leq \left(\frac{\hat{C}}{\tilde{C}+1} \right)^n [F(u^0) - F(u)], \\ \|u^n - u\|^p &\leq \frac{\hat{C}+1}{\tilde{C}} \left(\frac{\hat{C}}{\tilde{C}+1} \right)^n [F(u^0) - F(u)]. \end{aligned} \quad (25)$$

(ii) if $p > q$ we have

$$\begin{aligned} F(u^n) - F(u) &\leq \frac{F(u^0) - F(u)}{\left[1 + n \tilde{C} (F(u^0) - F(u))^{\frac{p-q}{q-1}} \right]^{\frac{q-1}{p-q}}}, \\ \|u - u^n\|^p &\leq \frac{(F(u^0) - F(u))^{\frac{q-1}{p-q}}}{\left[1 + (n-1) \tilde{C} (F(u^0) - F(u))^{\frac{p-q}{q-1}} \right]^{\frac{(q-1)^2}{(p-1)(p-q)}}}. \end{aligned} \quad (26)$$

The constants \hat{C} , \bar{C} and \tilde{C} are given in (28), (31) and (33), respectively.

Proof. From (21), using λ_M in (23), β_M in (22), and (24) in which we take $v_i = u_i$, $v = u$, $w = u^n$ and $w_i = w_i^{n+1}$, we have $F(u^{n+1}) - F(u) + \frac{A_M}{p} \|u - u^{n+1}\|^p \leq B_M \sum_{i=1}^m \|w_j^{n+1}\|^{q-1} \sum_{i=1}^m \|u_i - w_i^{n+1}\| \leq B_M m^{2-\frac{q}{p}} (\sum_{i=1}^m \|w_i^{n+1}\|^p)^{\frac{q-1}{p}} [(\sum_{i=1}^m \|w_i^{n+1}\|^p)^{\frac{1}{p}} + (\sum_{i=1}^m \|u_i\|^p)^{\frac{1}{p}}] \leq B_M m^{2-\frac{q}{p}} (\sum_{i=1}^m \|w_i^{n+1}\|^p)^{\frac{q-1}{p}} [(1+C_0)(\sum_{i=1}^m \|w_i^{n+1}\|^p)^{\frac{1}{p}} + C_0 \|u - u^n\|].$ Therefore, using (13) with $v = u^n$, (19), and λ_M given in (23), we have $F(u^{n+1}) - F(u) + \frac{A_M}{p} \|u - u^{n+1}\|^p \leq B_M (\frac{p}{A_M})^{\frac{q}{p}} m^{2-\frac{q}{p}} (F(u^n) - F(u^{n+1}))^{\frac{q-1}{p}} [(1+C_0)(F(u^n) - F(u^{n+1}))^{\frac{1}{p}} + C_0 (F(u^n) - F(u))^{\frac{1}{p}}] \leq B_M (\frac{p}{A_M})^{\frac{q}{p}} m^{2-\frac{q}{p}} (F(u^n) - F(u^{n+1}))^{\frac{q-1}{p}} [(1+2C_0)(F(u^n) - F(u^{n+1}))^{\frac{1}{p}} + C_0 (F(u^{n+1}) - F(u))^{\frac{1}{p}}].$ But, for some given $\eta > 0$ and $\zeta > 0$, we have $\zeta x^{\frac{1}{p}} - \eta x \leq (\frac{\zeta^p}{\eta})^{\frac{1}{p-1}}$, for any $x \geq 0$. Consequently, for a $0 < \eta < 1$, subtracting $\eta(F(u^{n+1}) - F(u))$ from both sides of the last inequality, we get

$$\begin{aligned} F(u^{n+1}) - F(u) + \frac{A_M}{p(1-\eta)} \|u - u^{n+1}\|^p &\leq \\ \hat{C} [F(u^n) - F(u^{n+1})]^{\frac{q-1}{p-1}}, \end{aligned} \quad (27)$$

where

$$\begin{aligned} \hat{C} = B_M (\frac{p}{A_M})^{\frac{q}{p}} m^{2-\frac{q}{p}} \left[(1+2C_0) (F(u^0) - F(u))^{\frac{p-q}{p(p-1)}} + \right. \\ \left. \left(B_M (\frac{p}{A_M})^{\frac{q}{p}} m^{2-\frac{q}{p}} \right)^{\frac{1}{p-1}} C_0^{\frac{p}{p-1}} / \eta^{\frac{1}{p-1}} \right] / (1-\eta), \end{aligned} \quad (28)$$

and we have used (17) to write $F(u^n) - F(u^{n+1}) \leq F(u^0) - F(u)$. From (27) we have

$$[F(u^{n+1}) - F(u)] \leq \hat{C} [F(u^n) - F(u^{n+1})]^{\frac{q-1}{p-1}}. \quad (29)$$

Using again (17) we have $F(u^n) - F(u^{n+1}) \leq F(u^n) - F(u)$, and from (13) and (23) we get $\frac{A_M}{p} \|u^{n+1} - u\|^p \leq F(u^{n+1}) - F(u)$. From these two last inequalities and (27) we get

$$\|u - u^{n+1}\|^p \leq \frac{\hat{C}}{C} [F(u^n) - F(u)]^{\frac{q-1}{p-1}}, \quad (30)$$

where

$$\bar{C} = \frac{(2-\eta) A_M}{(1-\eta)p}. \quad (31)$$

Now, if $p = q$, we can easily find (25) from (29) and (30). If $p \neq q$, we get from (29) that $F(u^{n+1}) - F(u) + \frac{1}{\hat{C}^{\frac{p-1}{q-1}}} [F(u^{n+1}) - F(u)]^{\frac{p-1}{q-1}} \leq$

$F(u^n) - F(u)$, and applying Lemma 3.2 in [9] we get $F(u^{n+1}) - F(u) \leq [\tilde{C} + (F(u^n) - F(u))^{\frac{q-1}{q-1}}]^{\frac{q-1}{q-p}}$, or

$$F(u^{n+1}) - F(u) \leq \left[(n+1)\tilde{C} + (F(u^0) - F(u))^{\frac{q-1}{q-1}} \right]^{\frac{q-1}{q-p}}, \quad (32)$$

where

$$\tilde{C} = \frac{p-q}{(p-1)(F(u^0) - F(u))^{\frac{p-q}{q-1}} + (q-1)\hat{C}^{\frac{p-1}{q-1}}}. \quad (33)$$

Equation (32) is another form of the first estimate in (26), and the second one can be obtained using (32) and (30). The value of η in the expression of \hat{C} and \tilde{C} can be arbitrary in $(0, 1)$. On the other hand, we see that the constants in the error estimations of $F(u^n) - F(u)$ in (25) and (26) are some increasing functions of \hat{C} , and there is an $\eta_0 \in (0, 1)$ such that $\hat{C}(\eta_0) \leq \hat{C}(\eta)$ for any $\eta \in (0, 1)$. This value η_0 can be found by solving a nonlinear algebraic equation. ■

4. The multiplicative Schwarz method as a subspace correction method

We shall prove in the following that for the problems in which we seek for the solution in a Sobolev space, Assumption 3.1 holds, and consequently, the convergence and error estimation theorems hold, too.

Let Ω be an open bounded domain in \mathbf{R}^d with Lipschitz continuous boundary $\partial\Omega$. We take $V = W_0^{1,s}(\Omega)$, $1 < s < \infty$, and a convex closed set $K \subset V$ satisfying

Property 4.1. *If $v, w \in K$, and if $\theta \in C^1(\Omega)$ with $0 \leq \theta \leq 1$, then $\theta v + (1-\theta)w \in K$.*

We consider an overlapping decomposition of the domain Ω ,

$$\Omega = \bigcup_{i=1}^m \Omega_i \quad (34)$$

in which Ω_i are open subdomains with Lipschitz continuous boundary. We associate to the domain decomposition (34) the subspaces $V_i = W_0^{1,s}(\Omega_i)$, $i = 1, \dots, m$. In this case, Algorithm 2.1 represents a multiplicative Schwarz method.

Remark 4.1 The above spaces V and V_i correspond to Dirichlet boundary conditions. Similar results can be obtained if we consider mixed boundary conditions. We take $\partial\Omega = \bar{\Gamma}_1 \cup \bar{\Gamma}_2$, $\Gamma_1 \cap \Gamma_2 = \emptyset$ a partition of

the boundary such that $\text{meas}(\Gamma_1) > 0$, and we consider the Sobolev space $V = \{v \in W^{1,s}(\Omega) : v = 0 \text{ on } \Gamma_1\}$. The subspaces V_i will be defined in this case as $V_i = \{v_i \in W^{1,s}(\Omega) : v_i = 0 \text{ in } \Omega - \bar{\Omega}_i, v_i = 0 \text{ in } \partial\Omega_i \cap \Gamma_1\}$, $i = 1, \dots, m$.

Also, we have considered problems having the solution in $W^{1,s}(\Omega)$, but all the obtained results hold with $[W^{1,s}(\Omega)]^N$, $N \geq 2$, in the place of $W^{1,s}(\Omega)$.

Concerning the decomposition (34), we assume that there are some functions $\theta_j^i \in C^1(\bar{\Omega})$, $i = 1, \dots, m$, $j = i, \dots, m$ such that for any $i = 1, \dots, m$ we have

$$\text{supp}(\theta_j^i) \subset (\bar{\Omega}_j), \quad 0 \leq \theta_j^i \leq 1, \quad j = i, \dots, m, \quad \sum_{j=i}^m \theta_j^i \equiv 1 \text{ in } \bigcup_{j=i}^m \Omega_j. \quad (35)$$

This is a easy enough constraint on the domain decomposition (34). In [8] or [1], for instance, some conditions in which a domain decomposition satisfies (35) are given.

Proposition 4.1 *If the domain decomposition (34) satisfies (35), then Assumption 3.1 holds for any convex set K having Property 4.1.*

Proof. The proof is similar to that given in [2] and we only outline it. Let us consider $w \in K$, $w_i \in V_i$ such that $w + \sum_{j=1}^i w_j \in K$, $i = 1, \dots, m$, and let v be another element in K . First we define $v_1 = \theta_1^1(v - w) + (1 - \theta_1^1)w_1$ and we prove that: $v_1 \in V_1$, $w + v_1 \in K$, $v - v_1 + w_1 \in K$, $v - w - v_1 \in W_0^{1,s}(\bigcup_{i=2}^m \Omega_i)$, and $v - w - v_1 = 0$ in $\Omega \setminus \overline{\bigcup_{i=2}^m \Omega_i}$. For $2 \leq i \leq m-1$, taking $v_i = \theta_i^i(v - w - \sum_{j=1}^{i-1} v_j) + (1 - \theta_i^i)w_i$ we recursively prove that: $v_i \in V_i$, $v_i + w + \sum_{j=1}^{i-1} w_j \in K$, $v - \sum_{j=1}^i v_j + \sum_{j=1}^i w_j \in K$, $v - w - \sum_{j=1}^i v_j \in W_0^{1,s}(\bigcup_{j=i+1}^m \Omega_j)$, and $v - w - \sum_{j=1}^i v_j = 0$ in $\Omega \setminus \overline{\bigcup_{j=i+1}^m \Omega_j}$. Finally, defining $v_m = v - w - \sum_{j=1}^{m-1} v_j$, we get that (1) and (2) hold. Also, (24) in Assumption 3.1 holds, in which C_0 depends on the unity partitions (35), but it is independent of w , v , w_i and v_i . ■

5. Numerical example

For a domain $\Omega \subset \mathbf{R}^d$ and an $1 < s < \infty$, let $K \subset V \equiv W_0^{1,s}(\Omega)$ be a closed and convex set. Given an $f \in V' \equiv W^{-1,s'}(\Omega)$, $1/s + 1/s' = 1$, we consider the problem

$$u \in K : \int_{\Omega} |\nabla u|^{s-2} \nabla u \nabla (v - u) \geq f(v - u), \quad \text{for any } v \in K. \quad (36)$$

The solution $u \in K$ of the above inequality is also the solution of the minimization problem $F(u) = \min_{v \in K} F(v)$, where $F(v) = \frac{1}{s} \int_{\Omega} |\nabla v|^s - f(v)$. We know (see [7]) that if $1 < s \leq 2$, then there exist two positive constants α and β such that $\langle F'(v) - F'(u), v - u \rangle \geq \alpha \frac{\|v - u\|_{1,s}^2}{(\|v\|_{1,s} + \|u\|_{1,s})^{2-s}}$, $\beta \|v - u\|_{1,s}^{s-1} \geq \|F'(v) - F'(u)\|_{V'}$, for any $v, u \in W_0^{1,s}(\Omega)$. Consequently, the functions introduced in (22) can be written as $\alpha_M(\tau) = \frac{\alpha}{(2M)^{2-s}} \tau^2$, $\beta_M(\tau) = \beta \tau^{s-1}$, and therefore, $A_M = \frac{\alpha}{(2M)^{2-s}}$, $B_M = \beta$, $p = 2$ and $q = s$ in (22). If $s \geq 2$, then there exist two positive constants α and β such that (see [4]) $\langle F'(v) - F'(u), v - u \rangle \geq \alpha \|v - u\|_{1,s}^s$, $\beta (\|v\|_{1,s} + \|u\|_{1,s})^{s-2} \|v - u\|_{1,s} \geq \|F'(v) - F'(u)\|_{V'}$, for any $v, u \in W_0^{1,s}(\Omega)$. Therefore, for a given $M > 0$, we have $\alpha_M(\tau) = \alpha \tau^s$, $\beta_M(\tau) = \beta (2M)^{s-2} \tau$, and therefore, $A_M = \alpha$, $B_M = \beta (2M)^{s-2}$, $p = s$ and $q = 2$ in (22). We can conclude from the above comments that Algorithm 2.1 can be applied for the solving of problem (36) if the convex set K has Property 4.1. Naturally, the error estimation in Section 3 hold.

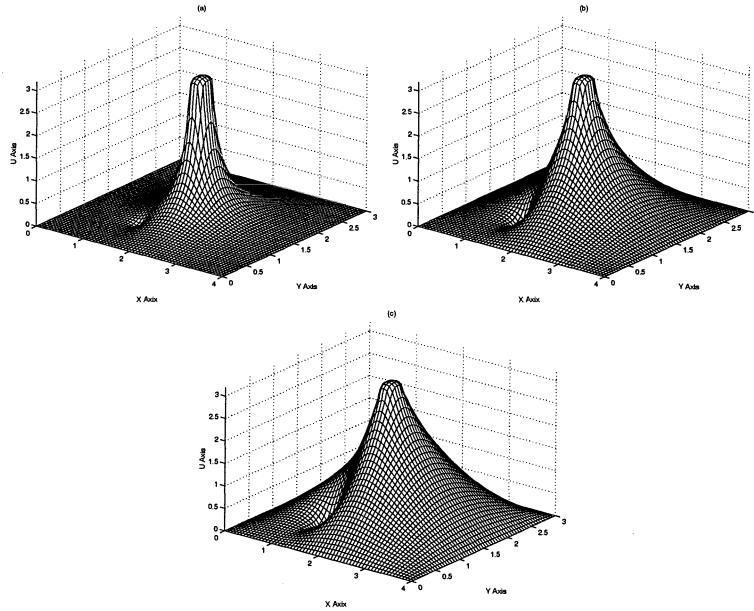


Figure 5.1. Solution for: (a) $s=1.5$, (b) $s=2.0$, (c) $s=3.0$.

If $\Omega \subset \mathbf{R}^2$ and the convex set is of the form $K = [a, b]$, where $a, b \in W_0^{1,s}(\Omega)$, $a \leq b$, then (36) is the problem of a nonlinear elastic membrane stretched over the obstacle a and under obstacle b . We have plotted in Figure 5.1 three computed solutions, corresponding to $s = 1.5$, $s = 2.0$ and $s = 3.0$, of such a problem having a rectangular domain. In this example, the exterior forces f are zero. In a subsequent paper we shall

present an analysis of the one and two-level Schwarz method in the finite element spaces, where the introduced assumption hold, too. In these cases, we are able to explicitly write the constant C_0 introduced in Assumption 3.1, as well as the constants in the error estimations of Theorem 2, as a function of the mesh and domain decomposition parameters.

Acknowledgments

The author acknowledges the financial support of IMAR under the contract nr. ICA1-CT-2000-70022 with the European Commission for this study.

References

- [1] L. Badea, A generalization of the Schwarz alternating method to an arbitrary number of subdomains, *Numer. Math.*, 55 (1989), pp. 61-81.
- [2] L. Badea, On the Schwarz alternating method with more than two subdomains for nonlinear monotone problems, *SIAM J. Numer. Anal.*, 28 (1991), pp. 179-204.
- [3] L. Badea, X.-C. Tai and J. Wang, Convergence rate analysis of a multiplicative Schwarz method for variational inequalities, *SIAM J. Numer. Anal.*, submitted, 2001.
- [4] Philippe G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
- [5] I. Ekeland and R. Temam, *Convex analysis and variational problems*, North-Holland, Amsterdam, 1976.
- [6] R. Glowinski, J. L. Lions and R. Trémolières, *Analyse numérique des inéquations variationnelles*, Dunod, 1976.
- [7] R. Glowinski and A. Marrocco, Sur l'approximation par éléments finis d'ordre un, et la résolution par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires, *Rev. Francaise Automat. Informat. Recherche Opérationnelle, Sér. Rouge Anal. Numér.*, R-2, 1975, pp. 41-76.
- [8] P. L. Lions, On the Schwarz alternating method I, in R. Glowinski et al., eds., First International Symposium on Domain Decomposition Methods for Partial Differential Equations, Philadelphia, *SIAM*, 1988, pp. 2-42.
- [9] X.-C. Tai and J. Xu, Global and uniform convergence of subspace correction methods for some convex optimization problems, *Math. of Comp.*, electronically published on 11 May, 2001.

NEW RESULTS ABOUT THE H-MEASURE OF A SET

Alina Bărbulescu

"Ovidius" University

Faculty of Mathematics and Informatics

8700 Constanta

Romania

abarbulescu@univ-ovidius.ro

Abstract The aim of this paper is to generalize our results ([1], [2]) related to the Hausdorff measure of a plane set.

Keywords: Hausdorff measure, δ – class Lipschitz function, equivalence.

1. Introduction

We denote by R^n the Euclidean n - dimensional space and by $d(E)$ - the diameter of a set $E \subset R^n$.

Definition 1.1 If $r_0 > 0$ is a fixed number, a continuous function $h(r)$, defined on $[0, r_0]$, nondecreasing and such that $\lim_{r \rightarrow 0} h(r) = 0$ is called a *measure function*. If $\delta \in R_+$, $E \subset R^n$ is a bounded set, the *Hausdorff h -measure of E* is defined by:

$$H_h(E) = \liminf_{\delta \rightarrow 0} \sum_i h(\rho_i)$$

inf being considered over all coverings of E with a countable number of spheres of radii $\rho_i \leq \delta$.

Definition 1.2 $f : D(\subset R^n) \rightarrow \overline{R}$ is a δ - class Lipschitz function if

$$|f(x + \alpha) - f(x)| \leq M |\alpha|^\delta, x \in D, \alpha \in R^n, x + \alpha \in D, M > 0. \quad (1)$$

Definition 1.3 Let $\varphi_1, \varphi_2 > 0$ be functions defined in a neighborhood of $0 \in R^n$. We say that φ_1 and φ_2 are equivalent and we denote by:

$\varphi_1 \sim \varphi_2$, for $x \rightarrow 0$, if there exist $r > 0$, $Q > 0$, satisfying:

$$\frac{1}{Q} \varphi_1(x) \leq \varphi_2(x) \leq Q \varphi_1(x), (\forall)x \in R^n, |x| < r. \quad (2)$$

An analogous definition can be given for $x \rightarrow \infty$. In this case, $\varphi_1 \sim \varphi_2$ means that the previous inequalities have place in all the space.

Definition 1.4 The graph of the function $f : [0, 1] \longrightarrow \overline{R}$ is the set:

$$\Gamma = \{(x, f(x)) | x \in [0, 1]\}.$$

2. Known results

Theorem 2.1 If h is a measure function such that

$$h(t) \sim t^p, p \geq 2, \quad (3)$$

and $f : [0, 1] \rightarrow \overline{R}$ is a δ -class Lipschitz function, with $\delta \in [0, 1]$, then: $H_h(\Gamma) < +\infty$. In the same hypothesis about h and f , the result remains true if $p \geq 1$ and $\delta > 1$.

Consider

$$g(x) = \begin{cases} 2x, & 0 \leq x < \frac{1}{2} \\ -2(x-1), & \frac{1}{2} \leq x < \frac{3}{2} \\ 2(x-2), & \frac{3}{2} \leq x < 2 \end{cases}$$

and:

$$f(x) = \sum_{i=1}^{\infty} \lambda_i^{-\delta} g(\lambda_i x), (\forall)x \in [0, 1], \quad (4)$$

where $\{\lambda_i\}_{i \in N^*}$ is a sequence of positive numbers.

Theorem 2.2

- a. If f is the function defined in (4), $\delta \in [0, 1]$, $\varepsilon > 1$, $\{\lambda_i\}_{i \in N^*}$ is a sequence of positive numbers, such that $\lambda_{i+1} > \varepsilon \lambda_i$, $(\forall) i \in N^*$ and h is a measure function, such that: $h(t) \sim t^p$, $p \geq 2$, then: $H_h(\Gamma) < +\infty$.
- b. In the same hypothesis about h , f and $\{\lambda_i\}_{i \in N^*}$, the result remains true if $\delta > 1$ and $\varepsilon > 1$.

3. New results

Theorem 3.1 *If h is a measure function such that*

$$h(t) \sim e^t t^p, p \geq 2, \quad (5)$$

and $f : [0, 1] \rightarrow \overline{R}$ is a δ - class Lipschitz function, with $\delta \in [0, 1]$, then: $H_h(\Gamma) < +\infty$. The result remains true if $p \geq 1$ and $\delta > 1$.

Proof. The first part of the proof follows that of [4].

First, we suppose that the coefficient in the Lipschitz inequality (1) can be taken 1, so that to any x corresponds an interval $(x - k, x + k)$ such that, for any $x + \alpha$ of this interval:

$$|f(x + \alpha) - f(x)| \leq |\alpha|^\delta.$$

Because $[0, 1]$ is a compact set, there exists a finite set of overlapping intervals covering $(0, 1)$:

$$(0, k_0), (x_1 - k_1, x_1 + k_1), \dots, (x_{n-1} - k_{n-1}, x_{n-1} + k_{n-1}), (1 - k_n, 1).$$

If c_i are arbitrary points, satisfying:

$$c_1 \in (0, x_1), c_i \in (x_{i-1}, x_i), i = 1, 2, \dots, n - 1, c_n \in (x_{n-1}, 1)$$

$$c_i \in (x_{i-1} - k_{i-1}, x_{i-1} + k_{i-1}) \cap (x_i - k_i, x_i + k_i), i = 1, \dots, n - 1.$$

we have: $0 < c_1 < x_1 < c_2 < x_2 < \dots < x_{n-1} < c_n < 1$.

The oscillation of $f(x)$ in the interval (c_{i-1}, c_i) is less than $2(c_i - c_{i-1})^\delta$ and thus the part of the curve corresponding to the interval (c_{i-1}, c_i) can be enclosed in a rectangle of height $2(c_i - c_{i-1})^\delta$ and of base $c_i - c_{i-1}$, and consequently in $[2(c_i - c_{i-1})^{\delta-1}] + 1$ squares of side $c_i - c_{i-1}$ or in the number of circles of radius $\frac{c_i - c_{i-1}}{\sqrt{2}}$ circumscribed about each of these squares.

We denoted by $[x]$ the integer part of x .

Given an arbitrary $r \in (0, \frac{1}{2})$ we can always assume: $c_i - c_{i-1} < r$, $i = 2, 3, \dots, n$.

Denote by C_r the set of all the above circles and consider

$$\sum_{C_r} h(2r) = \sum_{C_r} \left\{ \frac{h(2r)}{e^{2r}(2r)^p} \cdot e^{2r}(2r)^p \right\}. \quad (6)$$

$$r \in \left(0, \frac{1}{2}\right), e^{2r} \in (1, e) \quad (7)$$

We have to estimate $\sum_{C_r} (2r)^p$. The sum of the terms corresponding to the interval (c_{i-1}, c_i) is:

$$\begin{aligned}
 S &= \left\{ \left[2(c_i - c_{i-1})^{\delta-1} \right] + 1 \right\} \left\{ (c_i - c_{i-1}) \sqrt{2} \right\}^p \Leftrightarrow \\
 S &= 2^{\frac{p}{2}} (c_i - c_{i-1})^p \left\{ \left[2(c_i - c_{i-1})^{\delta-1} \right] + 1 \right\} . \\
 S &\leq 2^{\frac{p}{2}} (c_i - c_{i-1})^p \left\{ 2(c_i - c_{i-1})^{\delta-1} + 1 \right\} \Rightarrow \\
 S &\leq 2^{\frac{p}{2}+1} (c_i - c_{i-1})^{p+\delta-1} + 2^{\frac{p}{2}} (c_i - c_{i-1})^p \quad (8) \\
 c_i - c_{i-1} &< 1, p \geq 2, \delta \in [0, 1] \Rightarrow \\
 \left\{ \begin{array}{l} (c_i - c_{i-1})^p < c_i - c_{i-1} \\ p + \delta - 1 \geq 1 \Rightarrow (c_i - c_{i-1})^{p+\delta-1} < c_i - c_{i-1} \end{array} \right. \quad (9)
 \end{aligned}$$

From (8) and (9) it results:

$$\begin{aligned}
 S &\leq 2^{\frac{p}{2}+1} (c_i - c_{i-1}) + 2^{\frac{p}{2}} (c_i - c_{i-1}) = 3 \cdot 2^{\frac{p}{2}} (c_i - c_{i-1}) \Rightarrow \\
 \sum_{C_r} (2r)^p &\leq \sum_{i=2}^n 3 \cdot 2^{\frac{p}{2}} (c_i - c_{i-1}) = 3 \cdot 2^{\frac{p}{2}} \sum_{i=2}^n (c_i - c_{i-1}) \leq 3 \cdot 2^{\frac{p}{2}} \Leftrightarrow \\
 \sum_{C_r} (2r)^p &\leq 3 \cdot 2^{\frac{p}{2}} \quad (10)
 \end{aligned}$$

Using the definition 2 and the relations (7) and (10), (6) gives:

$$\sum_{C_r} h(2r) = \sum_{C_r} \left\{ \frac{h(2r)}{(2r)^p} (2r)^p \right\} < Qe \sum_{C_r} (2r)^p \leq 3 \cdot 2^{\frac{p}{2}} \cdot Q,$$

where $Q > 0$ and $r \in (0, \frac{1}{2})$, small enough.

Then $H_h(\Gamma) < +\infty$.

If $M \neq 1$, then $\sum_{C_r} h(2r) \leq 3 \cdot 2^{\frac{p}{2}} \cdot QM \Rightarrow H_h(\Gamma) < +\infty$.

If $p \geq 1$ and $\delta > 1$, then:

$$\begin{aligned}
 c_i - c_{i-1} &< 1, p \geq 1, \delta \geq 1 \Rightarrow \\
 \Rightarrow \left\{ \begin{array}{l} (c_i - c_{i-1})^p < c_i - c_{i-1} \\ p + \delta - 1 \geq 1 \Rightarrow (c_i - c_{i-1})^{p+\delta-1} < c_i - c_{i-1} \end{array} \right. \quad (11)
 \end{aligned}$$

and the proof is the same as above if we replace the relation (9) with (11). ■

Theorem 3.2 If h is a measure function such that

$$h(t) \sim P(t)e^{T(t)}, \quad (12)$$

where P and T are polynomials with positive coefficients:

$$\begin{aligned} P(t) &= a_1 t + a_2 t^2 + \dots + a_p t^p, \quad p \geq 1 \\ T(t) &= b_0 + b_1 t + \dots + a_m t^m, \end{aligned}$$

$f : [0, 1] \rightarrow \overline{\mathbb{R}}$ is a δ - class Lipschitz function, with $\delta \geq 1$, then: $H_h(\Gamma) < +\infty$. The result is also true if $p \geq 2$, $a_1 = 0$ and $\delta \in [0, 1]$.

Proof. The first part follows that of the previous theorem. We have to estimate the sum $\sum_{C_r} h(2r)$, for $r \in (0, \frac{1}{2})$.

$$\sum_{C_r} h(2r) = \sum_{C_r} \frac{h(2r)}{P(2r)e^{T(2r)}} \cdot P(2r)e^{T(2r)} < Q e^{\sum_{k=0}^m \frac{b_k}{2^k}} \sum_{C_r} P(2r), \quad (13)$$

because we have used the fact that $r \in (0, \frac{1}{2})$ and (12).

Now, we estimate $\sum_{C_r} P(2r)$. The sum of the terms corresponding to the interval (c_{i-1}, c_i) is:

$$S = \left\{ \left[2(c_i - c_{i-1})^{\delta-1} \right] + 1 \right\} \sum_{k=1}^p a_k \left((c_i - c_{i-1}) \sqrt{2} \right)^k,$$

where $[x]$ is the integer part of x .

$$\begin{aligned} S &\leq \left\{ 2(c_i - c_{i-1})^{\delta-1} + 1 \right\} \sum_{k=1}^p a_k (c_i - c_{i-1})^k 2^{k/2} \Leftrightarrow \\ S &\leq 2^{\frac{p}{2}} \max_{k=1,p} a_k \left\{ 2(c_i - c_{i-1})^{\delta-1} + 1 \right\} \sum_{k=1}^p (c_i - c_{i-1})^k \Leftrightarrow \\ S &\leq 2^{\frac{p}{2}} \max_{k=1,p} a_k \sum_{k=1}^p \left\{ 2(c_i - c_{i-1})^{k+\delta-1} + (c_i - c_{i-1})^k \right\}. \end{aligned} \quad (14)$$

If $p, \delta \geq 1$, then $k + \delta - 1 \geq 1$ and $(c_i - c_{i-1})^{k+\delta-1} \leq c_i - c_{i-1}$; thus:

$$S \leq 3 \cdot 2^{\frac{p}{2}} \max_{k=1,p} a_k \sum_{k=1}^p (c_i - c_{i-1})$$

and it results, from (13):

$$\sum_{C_r} h(2r) < Q e^{\sum_{k=0}^m \frac{b_k}{2^k}} \cdot 3 \cdot 2^{\frac{p}{2}},$$

where $Q > 0$ and $r \in (0, \frac{1}{2})$, small enough.

Then $H_h(\Gamma) < +\infty$.

If $p \geq 1, \delta \in [0, 1)$, then

$$1 \leq k - 1 \leq k + \delta - 1 < k$$

and

$$(c_i - c_{i-1})^k < (c_i - c_{i-1})^{k+\delta-1} \leq c_i - c_{i-1}.$$

Thus

$$S \leq 3 \cdot 2^{\frac{p}{2}} \max_{k=2,p} a_k \sum_{k=2}^p (c_i - c_{i-1})$$

and, analogous, it results that $H_h(\Gamma) < +\infty$. ■

Theorem 3.3

- a. If f is the function defined in (4), $\delta \in [0, 1], \varepsilon > 1, \{\lambda_i\}_{i \in N^*}$ is a sequence of positive numbers, such that $\lambda_{i+1} > \varepsilon \lambda_i$, $(\forall i \in N^*)$ and h is a measure function, satisfying the relation (5), then: $H_h(\Gamma) < +\infty$.
- b. In the same hypothesis about f and $\{\lambda_i\}_{i \in N^*}$, if $\delta > 1$ and $\varepsilon > 1$, then $H_h(\Gamma) < +\infty$.

Proof. The proof is analogous with that of the Theorem 2.2. (See [3].) ■

References

- [1] Barbulescu, A., *La finitude d'une h-mesure Hausdorff d'une ensemble de points dans le plan*, Analele Universitatii "Valahia" Târgoviste, 1995/1996, fasc.II, 93-99.
- [2] Barbulescu, A., *P-module and p-capacity*, Ph. D. Thesis, Iasi, 1997.
- [3] Barbulescu, A., *About the h - measure of a set*, Proceedings of the International Conference on Complex Analysis and Related Topics, Brasov, 2001 (to appear).
- [4] Besicovitch, A.S., Ursell, H. D., *Sets of fractional dimension (V): On dimensional numbers of some continuous curves*, London Math. Soc.J., **12** (1937), 118-125.

POSITIONAL MODELING IN A SYSTEM WITH TIME DELAY

Marina Blizorukova*

*Institute of Mathematics and Mechanics,
Ural Branch, Russian Academy of Sciences,
S. Kovalevskoi str., 16, Ekaterinburg, 620219 Russia
msb@imm.uran.ru*

Abstract The problem of dynamical modeling of unknown control of a system described by equations with time delay is discussed. The constructed algorithm operate synchronically with a process under consideration. It is stable with respect to informational noises and computational errors.

Keywords: modeling, time delay

1. Introduction. Statement of the problem

Let us consider a control system described by the vector equation with delay

$$\begin{aligned}\dot{x}(t) = & f_1(t, x(t), x(t - \alpha_1), \dots, x(t - \alpha_k)) + \\ & f_2(t, x(t), x(t - \alpha_1), \dots, x(t - \alpha_k))u(t),\end{aligned}\quad (1)$$

$$x_t(s) = x(t + s) = \varphi(s), \quad s \in [-\alpha, 0], \quad \varphi(\cdot) \in C_1[0, \alpha]. \quad (2)$$

Here $t \in T = [t_0, \vartheta]$, $t_0 < \vartheta$, is time, $x(t) \in R^n$ is the system's phase vector, $\alpha_i > 0$, $i = 1, \dots, k$ are constant delays, ($\alpha = \max\{\alpha_i : i = 1, 2, \dots, k\}$), $u \in R^p$ is a control, $f_1(\cdot)$, $f_2(\cdot)$ are continuous vector and matrix functions satisfying on $T \times R^n \times \dots \times R^n$ the Lipschitz condition:

$$\begin{aligned}|f_j(t_1, x_1, y_1^{(1)}, \dots, y_k^{(1)}) - f_j(t_2, x_2, y_1^{(2)}, \dots, y_k^{(2)})| \leq \\ c_j(|t_1 - t_2| + |x_1 - x_2| + \sum_{i=1}^k |y_i^{(1)} - y_i^{(2)}|), \quad c_j \in (0, +\infty)\end{aligned}$$

*This work was supported in part by the Russian Foundation for Basic Research (grant # 01-01-00566).

for $j = 1, 2, t_1, t_2 \in T$, $x_1, x_2, y_1^{(1)}, \dots, y_k^{(1)}, y_1^{(2)}, \dots, y_k^{(2)} \in R^n$. Throughout the paper we denote by $|\cdot|$ the Euclidean norm in corresponding spaces R^n or $R^{n \times p}$, $C_1[0, \alpha]$ is the space of continuously differentiable functions; $x_t(s) = x(t_0 + s)$, $s \in [-\alpha, 0]$.

Further, any measurable (by Lebesgue) function $u(\cdot)$ from a fixed compact $P \subset R^p$ is called a control. A solution $x(\cdot)$ of system (1) (in the sense of Caratheodory) with the initial state (2) is called a motion of the system generated by control $u(\cdot)$.

The problem under consideration is in the following. We have the system (1) under the action of some control $u(\cdot)$, $u(t) \in P$ for a.e. $t \in T$. This control $u(\cdot)$ as well as the trajectory $x(\cdot)$ generated by $u(\cdot)$ are unknown. During the process, at frequent enough instances $t = \tau_i$ the components of state vector $x(\tau_i)$ are measured. The results of inaccurate measurements are vectors ξ_i satisfying the inequalities $|\xi_i - x(\tau_i)| \leq \nu_i$, where ν_i is an observation error at the moment τ_i . The problem consists in construction of an algorithm that calculates (inaccurately) in real time mode some function $v_h(\cdot)$ from the set $U(x(\cdot))$ of controls generating $x(\cdot)$

$$U(x(\cdot)) = \{u(\cdot) \in P : x(t) = x(t; t_0, x_t(\cdot), u(\cdot)) \text{ a.e. } t \in T\}$$

One of approaches to solving similar problems was presented and developed in [1–3]. For delay systems this method was studied in [4, 5]. In this work, an algorithm of dynamical reconstruction of an input in a system with delay different from [4, 5] is suggested. This algorithm is based on the constructions from [2]. It should be noted that in all papers mentioned above the results of inaccurate measurements satisfy the inequalities $|\xi_i - x(\tau_i)| \leq h$, $h \in (0, 1)$.

2. The solving algorithm

We search for the solution of the problem in the class of finite-step dynamical algorithms, i. e. such algorithms that use information only in the nodes of a finite partition and process it between the nodes. As a solution method we choose the method of dynamical reconstruction of controls combined with the well-known in the theory of ill-posed problems discrepancy method [7]. Briefly, the latter consists in the following. Some set containing the sought element is defined on the basis of available inaccurate information. After this, another element, which approximates the one mentioned above (usually, this element is sought as an extremum of a suitable functional) is indicated by some rule.

Let $x_*(\cdot)$ be a fixed real motion of the system. We denote by $u_*(\cdot)$ the element of the set $U(x(\cdot))$ of minimal $L_2(T; U)$ -norm. Let for every $h \in (0, 1)$ a uniform partition of the segment T

$$\Delta_h = \{\tau_{h,i}\}_{i=0}^{m_h}, \quad t_0 = \tau_{h,0} < \tau_{h,1} < \dots < \tau_{h,m_h} = \vartheta, \quad \tau_{h,i} = \tau_{h,i-1} + \delta(h)$$

with the property

$$h/\delta(h) \rightarrow 0, \quad \delta(h) \rightarrow 0, \quad \text{as } h \rightarrow 0$$

be fixed. Let the motion $x(t)$ and the values ξ_i^h :

$$|\xi_i^h - x(\tau_{h,i})| \leq \nu_i^h$$

belong to an a priori given compact set $E \subset R^n$. We assume that observation errors $\nu_i^h \geq 0$ satisfy the property

$$\varphi(h) = \sum_{i=1}^{m_h} \nu_i^h \rightarrow 0 \quad \text{as } h \rightarrow 0.$$

Latter condition may be meaningfully treated, for example, in the following sense. The closer is the value h to zero, the better the set of results of measurements of $x(\cdot)$ approximates in average on T this trajectory, i. e.

$$\sum_{i=0}^{m_h} \delta(h) |\xi_i^h - x(\tau_{h,i})| \leq \delta(h) \varphi(h) \rightarrow 0 \quad \text{as } h \rightarrow 0.$$

Introduce constants $c_0 > 0$, $c_u > 0$, $c_* > 0$ such that

$$\begin{aligned} |f_1(t, x, y_1, \dots, y_k) + f_2(t, x, y_1, \dots, y_k)u| &\leq c_0 \quad \forall t \in T, \\ x, y_i \in E, \quad i = 1, 2, \dots, k, \quad u \in P; \quad |u| &\leq c_u \quad \forall u \in P; \\ |\varphi(t_1) - \varphi(t_2)| &\leq c_0 |t_1 - t_2| \quad \forall t \in [0, \alpha]; \\ |x(\tau_i)| &\leq c_*, \quad |\xi_i^h| \leq c_*, \quad \forall i \in [0 : m_h], \quad h \in (0, 1). \end{aligned} \tag{3}$$

Let us turn to the description of the algorithm. We choose an auxiliary discrete control system (a model) corresponding to system (1).

The model state at every moment $\tau_{h,i}$ is characterized by the vector

$$\begin{aligned} z^h(\tau_{h,i+1}) &= z^h(\tau_{h,i}) + \\ &[f_1(\tau_{h,i-1}, \xi_{i-1}^h, \xi_h(\tau_{h,i-1} - \alpha_1), \dots, \xi_h(\tau_{h,i-1} - \alpha_k)) + \\ &f_2(\tau_{h,i-1}, \xi_{i-1}^h, \xi_h(\tau_{h,i-1} - \alpha_1), \dots, \xi_h(\tau_{h,i-1} - \alpha_k))v_h(\tau_{h,i})]\delta, \end{aligned} \tag{4}$$

$$z_h(\tau_{h,1}) = \varphi(0), \tag{5}$$

where

$$\xi_h(\tau_{h,i} - \alpha_j) = \xi_{i-m_h^0(j)}^h,$$

$$\vartheta_{m_h(j)} = \alpha_j - m_h(j)\delta(h), \quad m_h^0(j) = \begin{cases} m_h(j), & \vartheta_{m_h(j)} = 0, \\ m_h(j) + 1, & \vartheta_{m_h(j)} \neq 0. \end{cases}$$

The symbol $m_h(j)$ denotes the integral part of the number $\alpha_j/\delta(h)$.

A model control is chosen according to the principle of extremal shift [6]. We suppose that at every moment $\tau_{h,i}$ ($i \geq 1$) the closed set

$$\begin{aligned} \Omega_{h,i} = \Omega_{h,i}(\xi_i^h, \xi_{i-1}^h) &= \{v \in P : (z_h(\tau_{h,i}) - \xi_{i-1}^h)' \times \\ &\left[f_1(\tau_{h,i-1}, \xi_{i-1}^h, \xi_h(\tau_{h,i-1} - \alpha_1), \dots, \xi_h(\tau_{h,i-1} - \alpha_k)) + \right. \\ &\left. f_2(\tau_{h,i-1}, \xi_{i-1}^h, \xi_h(\tau_{h,i-1} - \alpha_1), \dots, \xi_h(\tau_{h,i-1} - \alpha_k))v \right] - \\ &(z_h(\tau_{h,i}) - \xi_i^h)' \frac{\xi_i^h - \xi_{i-1}^h}{\delta} \leq \sigma_{h,i}^\delta \} \end{aligned} \quad (6)$$

is constructed on the basis of measurement results ξ_i^h and ξ_{i-1}^h . Here the prime stands for transposition, the values $\sigma_{h,i}^\delta$ being of the form

$$\begin{aligned} \sigma_{h,i}^\delta &= c_0^2 \delta + c_0(\nu_i^h + \nu_{i-1}^h) + (2c_* + \nu_{i-1}^h)(c_1 + c_2 c_u)[c_0(k+2)\delta + \nu_{i-1}^h + \\ &\sum_{j=1}^k (\nu_{i-m_h(j)}^h + \nu_{i-1-m_h(j)}^h) + \frac{2c_*}{\delta}(\nu_i^h + \nu_{i-1}^h)]. \end{aligned}$$

$\nu_{i-m_h(j)}^h = 0$ for $i - m_h(j) < 0$. Following the ideology of discrepancy method, we indicate the law of forming the control in the model:

$$v_h(\tau_{h,i}) = \begin{cases} \arg \min \{|u| : u \in \Omega_{h,i}\}, & \text{if } \Omega_{h,i} \neq \emptyset, \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

Introduce the family of functions

$$v_h(t) = v_h(\tau_{h,i}) \quad \text{for } t \in [\tau_{h,i}, \tau_{h,i+1}), \quad i \in [0, m_h - 1]. \quad (8)$$

The following theorem is true.

Theorem 1 *A family of controls $\{v_h(\cdot)\}$, $h \in (0, 1)$, defined by (4)–(8), satisfies the following property*

$$|v_h(\cdot) - u_*(\cdot)|_{L_2(T; U)} \rightarrow 0 \quad \text{as } h \rightarrow 0. \quad (9)$$

The proof of this theorem is based on the following auxiliary statement.

We introduce a value $\varepsilon_{j+1} = |z_{j+1} - x_j|^2$, $x_j = x_*(\tau_j)$, $z_j = z_h(\tau_j)$.

Lemma 1 *The inequality*

$$\varepsilon_j \leq 2\delta(h) \sum_{i=0}^j \sigma_{n,i}^\delta + 6c_0\delta(\vartheta - t_0) + 4(c_* + c_0) \sum_{i=0}^j \nu_i^h$$

is true.

Proof of Lemma 1. For $\vartheta_{m_h(j)} = 0$, we get

$$|x(t - \tau_j) - \xi_{i-1,j}^h|_n \leq c_0\delta + (\nu_{i-m_h(j)}^h + \nu_{i-1-m_h(j)}^h), \quad j \in [1 : k],$$

where $t \in [t_{i-1}, t_i]$. If $\vartheta_{m_h(j)} \neq 0$, then $|x(t) - \xi_{i-1}^h|_n \leq c_0\delta + \nu_{i-1}^h$. Therefore,

$$|x(t) - \xi_{i-1}^h|_n + |x(t - \tau_1) - \xi_{i-1,1}^h|_n + \dots + |x(t - \tau_k) - \xi_{i-1,k}^h|_n \leq \\ c_0\delta(k+1) + \nu_{i-1}^h + \sum_{j=1}^k (\nu_{i-m_h(j)}^h + \nu_{i-1-m_h(j)}^h).$$

It is easy to see that (for $c_0 \geq 1$)

$$|\Delta f_1|_n \leq c_1 \left(\delta c_0(k+2) + \nu_{i-1}^h + \sum_{j=1}^k (\nu_{i-m_h(j)}^h + \nu_{i-1-m_h(j)}^h) \right).$$

We obtain analogously

$$|\Delta f_2|_n \leq c_u c_2 \left(\delta c_0(k+2) + \nu_{i-1}^h + \sum_{j=1}^k (\nu_{i-m_h(j)}^h + \nu_{i-1-m_h(j)}^h) \right).$$

Hence,

$$|\Delta f|_n \leq \mu(\delta, i) = (c_1 + c_2 c_u) c_0(k+2)\delta + \\ + (c_1 + c_2 c_u) \left(\nu_{i-1}^h + \sum_{j=1}^k (\nu_{i-m_h(j)}^h + \nu_{i-1-m_h(j)}^h) \right). \quad (10)$$

Taking into account the following inequalities

$$|z^h(t_i) - \xi_{i-1}^h|_n \leq |z^h(t_i)|_n + |\xi_{i-1}^h - x(t_{i-1})|_n + |x(t_{i-1})|_n \leq \\ \leq 2c_* + \nu_{i-1}^h, \\ |\xi_i^h - \xi_{i-1}^h|_n \leq |\xi_i^h - x(t_i)|_n + |\xi_{i-1}^h - x(t_{i-1})|_n + |x(t_i) - x(t_{i-1})|_n \leq c_0\delta + \nu_i^h + \nu_{i-1}^h, \quad (11)$$

we have

$$\mu_i^{(2)} = (\xi_i^h - \xi_{i-1}^h, \int_{t_{i-1}}^{t_i} [f_1(t, x(t), x(t - \tau_1), \dots, x(t - \tau_k)) + \\ + f_2(t, x(t), x(t - \tau_1), \dots, x(t - \tau_k)) u_*(t)] dt)_n \leq \\ |\xi_i^h - \xi_{i-1}^h|_n \int_{t_{i-1}}^{t_i} c_0 dt \leq c_0\delta(c_0\delta + \nu_i^h + \nu_{i-1}^h).$$

From (10), (11) we also deduce

$$\mu_i^{(1)} \leq |z^h(t_i) - \xi_{i-1}^h|_n \delta |\Delta f|_n \leq (2c_* + h_{i-1}) \delta \mu(\delta, i).$$

This and the assumption that $\delta \in (0, 1)$, $h \in (0, 1)$, $h/\delta \in (0, 1)$ imply

$$\mu_i \leq \rho_\delta(i)\delta, \quad (12)$$

where $\rho_\delta(i) = c_0(c_0\delta + \nu_i^h + \nu_{i-1}^h) + (2c_* + \nu_{i-1}^h)\mu(\delta, i)$.

Using (11), (12), we get

$$\begin{aligned} & (z^h(t_i) - \xi_{i-1}^h, f_1(t_{i-1}, \xi_{i-1}^h, \xi_{i-1,1}^h, \dots, \xi_{i-1,k}^h) + \\ & f_2(t_{i-1}, \xi_{i-1}^h, \xi_{i-1,1}^h, \dots, \xi_{i-1,k}^h)\delta^{-1} \int_{t_{i-1}}^{t_i} u_*(\tau) d\tau)_n - \\ & (z^h(t_i) - \xi_i^h, \xi_i^h - \xi_{i-1}^h)_n \delta^{-1} \leq \rho_\delta(i) + (2c_* + \nu_{i-1}^h)(\nu_i^h + \nu_{i-1}^h)\delta^{-1}. \end{aligned} \quad (13)$$

Further,

$$\begin{aligned} \varepsilon_{j+1} & \leq \varepsilon_j + 2(z_j - \xi_{j-1}^h, [f_1(t_{j-1}, \xi_{j-1}^h, \xi_{j-1,1}^h, \dots, \xi_{j-1,k}^h) + \\ & f_2(t_{j-1}, \xi_{j-1}^h, \xi_{j-1,1}^h, \dots, \xi_{j-1,k}^h)v^h(t_j)])_n \delta + 2\nu_{j-1}^h c_0 \delta + \\ & 2(z_j - x_j, x_j - x_{j-1})_n + 4(c_0 \delta)^2 + 2|x_j - x_{j-1}|_n^2. \end{aligned}$$

Hence,

$$\begin{aligned} \varepsilon_{j+1} & \leq \varepsilon_j + 2(z_j - \xi_{j-1}^h, [f_1(t_{j-1}, \xi_{j-1}^h, \xi_{j-1,1}^h, \dots, \xi_{j-1,k}^h) + \\ & f_2(t_{j-1}, \xi_{j-1}^h, \xi_{j-1,1}^h, \dots, \xi_{j-1,k}^h)v^h(t_j)])_n \delta + \\ & 2(z_j - \xi_j^h, x_j - x_{j+1})_n + 6(c_0 \delta)^2 + 2c_0 \delta (\nu_j^h + \nu_{j-1}^h). \end{aligned}$$

Besides,

$$\begin{aligned} |(z_j - \xi_j^h, x_j - x_{j-1})_n - (z_j - \xi_j^h, \xi_j^h - \xi_{j-1}^h)_n| & \leq \\ & (\nu_j^h + \nu_{j-1}^h)|z_j - \xi_j^h|_n \leq 2c_*(\nu_j^h + \nu_{j-1}^h). \end{aligned}$$

Therefore

$$\varepsilon_{j+1} \leq \varepsilon_j + 2\sigma_{h,j}^\delta \delta + 2(c_* + c_0 \delta)(\nu_j^h + \nu_{j-1}^h) + 6(c_0 \delta)^2.$$

Finally, we have

$$\varepsilon_{j+1} \leq 2\delta \sum_{i=0}^j \sigma_{h,i}^\delta + 6c_0^2 \delta (\vartheta - t_0) + 2(c_* + c_0 \delta) \sum_{i=1}^j (\nu_i^h + \nu_{i-1}^h).$$

This finishes the proof of the lemma.

Remark In the case when the set P is a ball:

$$P = \{u \in R^N : |u|_N \leq a\},$$

the control in the model can be written explicitly:

$$v_i^h = \begin{cases} 0, & \text{if } b_i^h = 0 \text{ or } \chi_i^h > 0 \\ \chi_i^h \frac{b_i^h}{|b_i^h|_n^2}, & \text{if } b_i^h \neq 0 \text{ and } \chi_i^h \leq 0. \end{cases}$$

Here

$$\begin{aligned} b_i^h &= (z_i - \xi_{i-1}^h, f_2(t_{i-1}, \xi_{i-1}^h, \xi_{i-1,1}^h, \dots, \xi_{i-1,k}^h))_n, \\ \chi_i^h &= \sigma_h^\delta + (z_i - \xi_{i-1}^h, \frac{\xi_i^h - \xi_{i-1}^h}{\delta} - f_1(t_{i-1}, \xi_{i-1}^h, \xi_{i-1,1}^h, \dots, \xi_{i-1,k}^h))_n. \end{aligned}$$

3. Conclusion

The problem of dynamical modeling (reconstruction) of an unknown input that determines the motion of a dynamical system with time delay through given inaccurate measurements of the system's current state is considered. The scheme of solving this problem leans upon the method of auxiliary positionally controlled models. The suggested solution method consists in the combination of the method of dynamical reconstruction of controls and a dynamical modification of the well-known in the theory of ill-posed problems discrepancy method. A dynamical algorithm working "in real time" based on this method is discussed.

References

- [1] Kryazimskii, A. V. and Osipov, Yu. S. (1983). Modelling of a control in a dynamic system. *Engineering Cybernetics*, 21:38–47. in Russian.
- [2] Kryazimskii, A. V. and Osipov, Yu. S. (1988). *On methods of positional modeling of a control in a dynamic system*. 34–44. Qualitative questions of the theory of differential equations and control systems, Urals. Sci. Center, Sverdlovsk. in Russian.
- [3] Kryazimskii, A. V. and Osipov, Yu. S. (1995). *Inverse Problems for Ordinary Differential Equations: Dynamical Solutions*. Gordon and Breach. London.
- [4] Kryazimskii, A. V., Maksimov, V. I. and Osipov, Yu. S. (1983). On positional modelling in dynamic systems. *Prikl. math. mech.*, 47:883–889.
- [5] Blizorukova, M.S. and Maksimov, V.I. (1999). On the reconstruction of a pair "control–trajectory" in a system with hereditary. *Problems of control and informatics*, 4:37–48. in Russian.
- [6] Krasovskii, N.N. and Subbotin, A.I. (1988). *Game-Theoretical Control Problems*. Springer Verlag. New York – Berlin.
- [7] Tikhonov, A.N. and Arsenin, V.Ya. (1977). *Solution of Ill-posed Problems*. John Wiley, New York.

UNIFORM STABILITY OF A COUPLED SYSTEM OF HYPERBOLIC/PARABOLIC PDE'S WITH INTERNAL DISSIPATION*

Francesca Bucci

Università degli Studi di Firenze

Dipartimento di Matematica Applicata "G. Sansone"

Via S. Marta 3, I-50139 Firenze, ITALY

fbucci@dma.unifi.it

Abstract We study the uniform stability of a coupled system of hyperbolic/parabolic partial differential equations (PDEs) with nonlinear internal dissipation. We analyze both the case of distributed damping on the entire domain, and the case of damping with localised support. In the corresponding stability results, decay rates of weak solutions to the PDE system under consideration are described via the solutions to appropriate nonlinear ordinary differential equations.

Keywords: coupled partial differential equations, uniform decay rates, saturation, locally distributed damping, multipliers' method.

Introduction

The problem of stabilization of coupled or interconnected systems of Partial Differential Equations (PDEs) has become in the last several years a central topic of mathematical control theory of infinite-dimensional systems [11, 9]. The present paper is focused on uniform stabilization of a coupled system of hyperbolic/parabolic PDEs, which is a nonlinear generalization of a PDE model originated in [6].

The mathematical model studied in [6] consists of a wave equation in a bounded domain Ω of \mathbb{R}^3 (the "acoustic chamber"), which is strongly coupled to a linear (abstract) structurally damped plate-like equation acting only on the elastic, flat wall of the chamber (the *interface*). A distinguished feature of this model is that it displays a boundary dissipation.

*Research supported by the Italian Ministero dell'Istruzione, dell'Università e della Ricerca.

pation term of the wave component which accounts for lack of uniform stability of the overall system, even in the presence of viscous damping on the entire domain (*overdamping* phenomenon). In [6] it has been shown that by introducing a comparable static damping in the boundary condition of the wave component, then the corresponding feedback system is (exponentially) uniformly stable.

In this paper we study the stability properties of that ultimate model when internal damping is subject to *nonlinear* effects, namely system (1.1) described in the next section. More precisely, we aim to establish uniform decay rates for the (natural) energy $E(t)$ of the corresponding weak solutions, as $t \rightarrow +\infty$. We shall examine both (i) the case of (nonlinear) damping distributed on the entire domain, and (ii) the more challenging case when internal damping is active in a thin layer near the interface.

In the former situation uniform decay rates of the underlying energy are obtained, without assuming *a priori* growth conditions on the nonlinear function F near the origin. Furthermore, we allow maximum polynomial growth of F at infinity (up to power five). Decay rates are described via the solution to an appropriate nonlinear ODE (see Theorem 2.2).

Unlike most literature, our analysis will include the case when the dissipation term is bounded at infinity (*saturation*). In this significant case uniform decay rates are achieved for the solutions with initial data which belong to a slightly smoother function space than the energy space [4]. We stress that the novelty of this second result is that

- we obtain uniform decay rates of the energy of solutions corresponding to a bit smoother initial data, rather than of *strong* solutions (i.e. solutions corresponding to initial data in the domain of the dynamic operator), as done, e.g., in [14] in the case of dissipative wave equations;
- in addition, we still do not require any *a priori* assumption on the growth of F near the origin.

Next, if appropriate geometric conditions on Ω are in force (Hypothesis 3.1), by restricting the polynomial growth at infinity on the nonlinearity, we are able to show that a locally distributed damping near the interface is in fact sufficient to stabilize the system at a uniform decay rate (Theorem 3.6), thus completing the analysis of case (ii).

We note that these stability results obtained for the coupled PDE system (1.1) hold, as well, for the single, uncoupled wave equation (even without “overdamping” term). A technical comparison with the vast

literature on stabilization of dissipative wave equations can be found in [4].

As a final comment we point out that it would be very interesting to study (infinite-dimensional) Hamilton-Jacobi-Bellman (HJB) Equations associated with optimal control problems for abstract equations in Banach spaces of the form $y' = Ay + Bu + F(y)$, when (a) the control operator B is *unbounded*, and (b) the operator $e^{At}B$ satisfies the so called “singular estimate” (*cf.* [1, 9, 12, 6]). Let us recall that in the special Linear Quadratic (LQ) case, recently it has been shown that in the framework defined by (a)-(b), namely if the singular estimate holds, yet in the absence of analyticity of the underlying semigroup, then well-posedness of associated Riccati equations, along with the sought-after property that the gain operator is *bounded*, follow as well [12]. The question whether in the more general case of control problems for *semi-linear* equations subject to (a) and (b) (and with possibly non quadratic cost functionals) one can develop a theory of HJB equations closely akin to the one relative to the “analytic” class of systems with unbounded control operator—such as it arises in the context of nonlinear control problems for a parabolic PDE with boundary or point control [7, 8]—is still an open problem.

1. The PDE model

The PDE model under consideration in this paper is a nonlinear generalization of a coupled system of hyperbolic/parabolic PDEs studied in [6]. A brief mathematical description of this system is given below; see [6] for a thorough analysis of the linear problem.

Let $\Omega \subset \mathbb{R}^3$ be an open bounded domain with boundary $\Gamma = \overline{\Gamma_0 \cup \Gamma_1}$, where Γ_0 and Γ_1 are open, connected, disjoint parts, $\Gamma_0 \cap \Gamma_1 = \emptyset$ in \mathbb{R}^2 , of positive measure. The sub-boundary Γ_0 is flat and is referred to as the elastic (or flexible) wall, while Γ_1 is referred to as the rigid (or hard) wall. At the outset, we assume that either Ω is sufficiently smooth (say, Γ is of class C^2), or else Ω is convex.

The acoustic medium in the chamber Ω is described by the wave equation in the variable z , while v represents the (abstract) deflection of the abstract *structurally damped* plate equation on Γ_0 . The interaction between wave and plate takes place on Γ_0 (the *interface*).

Thus, our goal is to investigate the stability properties of the following system of coupled PDEs:

$$\left\{ \begin{array}{ll} z_{tt} = \Delta z - F(z_t) & \text{in } Q = (0, \infty) \times \Omega \\ \frac{\partial z}{\partial \nu} + d_1 z = 0 & \text{on } \Sigma_1 = (0, \infty) \times \Gamma_1 \\ \frac{\partial z}{\partial \nu} + D_0 z_t + \beta D_0 z = v_t & \text{on } \Sigma_0 = (0, \infty) \times \Gamma_0 \\ z(0, \cdot) = z^0, \quad z_t(0, \cdot) = z^1 & \text{in } \Omega \\ v_{tt} + \mathcal{A}v + \rho \mathcal{A}^\alpha v_t + z_t|_{\Gamma_0} = 0 & \text{on } \Sigma_0 = (0, \infty) \times \Gamma_0 \\ v(0, \cdot) = v^0, \quad v_t(0, \cdot) = v^1 & \text{in } \Gamma_0. \end{array} \right. \quad (1.1)$$

Basic assumptions. The assumptions pertaining to the *linear* operators \mathcal{A} and D_0 in (1.1) will be held throughout the paper and will not be mentioned explicitly any more.

- (H0) \mathcal{A} (the elastic operator): $L^2(\Gamma_0) \supset \mathcal{D}(\mathcal{A}) \rightarrow L^2(\Gamma_0)$ is a positive, self-adjoint operator. Moreover, $\rho > 0$, $\beta \geq 0$, $d_1 > 0$ and $\frac{1}{2} \leq \alpha \leq 1$ are constants.
- (H1) $D_0 : L^2(\Gamma_0) \supset \mathcal{D}(D_0) \rightarrow L^2(\Gamma_0)$ is a positive, self-adjoint operator, and there exists a constant r_0 , $0 \leq r_0 \leq 1/4$, and positive constants δ_1 , δ_2 such that

$$\begin{aligned} \delta_1 \|z\|_{\mathcal{D}(\mathcal{A}^{r_0})}^2 &\leq (D_0 z, z)_{L^2(\Gamma_0)} \leq \delta_2 \|z\|_{\mathcal{D}(\mathcal{A}^{r_0})}^2 \\ \forall z \in \mathcal{D}(\mathcal{A}^{r_0}) \subset \mathcal{D}(D_0^{1/2}). \end{aligned} \quad (1.2)$$

Moreover, it is assumed that $H^1(\Gamma_0) \subseteq \mathcal{D}(D_0^{1/2})$; that is,

$$D_0^{1/2} : \text{continuous } H^1(\Gamma_0) \rightarrow L^2(\Gamma_0).$$

In particular, for simplicity of exposition, we shall examine only the case when D_0 is the realization of a differential operator of order s , with $1 < s \leq 2$. Accordingly, it is assumed that $\beta > 0$, see [6, Sections 1.1–1.2].

Instead, the mathematical features of the dissipation term $F(z_t)$ for the wave component will be made more precise in the different frameworks of model (1.1) that we are going to consider. We recall that a thorough analysis of stability properties of system (1.1) with *full viscous damping* (i.e. with $F(z_t) = z_t$), has been performed in [6]. Here, we shall examine first the case when (1.1) displays nonlinear damping distributed

on the entire domain Ω , next the more challenging case of damping with localised support. Specific assumptions on the nonlinear function F and possible geometric constraints on Ω will be introduced correspondingly.

Function spaces. Before going into the heart of the stability issue, we need to introduce the natural state space for system (1.1). It is known from [6] that the function space Y_1 which is needed to describe the wave component of system (1.1) is given by $Y_1 := Z \times L^2(\Omega)$, with

$$Z := \left\{ f \in H^1(\Omega) : f|_{\Gamma_0} \in \mathcal{D}(D_0^{1/2}) \right\}, \quad (1.3)$$

endowed with the norm

$$\|f\|_Z^2 := \|\nabla f\|_{L_2(\Omega)}^2 + \beta \|D_0^{1/2} f|_{\Gamma_0}\|_{L_2(\Gamma_0)}^2. \quad (1.4)$$

Then, the state space Y for problem (1.1) is given by

$$Y := Z \times L_2(\Omega) \times \mathcal{D}(\mathcal{A}^{1/2}) \times L^2(\Gamma_0). \quad (1.5)$$

Thus, the energy $E(t)$ of weak solutions $[\vec{z}(t, \cdot), \vec{v}(t, \cdot)]$ to the coupled system (1.1) is defined by

$$E(t) = E_z(t) + E_v(t); \quad (1.6a)$$

$$E_z(t) := \|\nabla z(t)\|_{L_2(\Omega)}^2 + \beta \|D_0^{1/2} z(t)\|_{L_2(\Gamma_0)}^2 + \|z_t(t)\|_{L_2(\Omega)}^2 \quad (1.6b)$$

$$E_v(t) := \|\mathcal{A}^{1/2} v(t)\|_{L_2(\Gamma_0)}^2 + \|v_t(t)\|_{L_2(\Gamma_0)}^2, \quad (1.6c)$$

where $E_z(t)$ and $E_v(t)$ denote the wave and plate energy, respectively.

Remark 1.1. We note that due to space constraints, we shall omit from this article the description of the abstract set-up for problem (1.1) and the analysis of well-posedness of the corresponding nonlinear evolution equation (see [6, 4]).

Notation. In order to simplify the notation, hereafter the norms $\|\cdot\|_{H^s(\Omega)}$ and $\|\cdot\|_{H^s(\Gamma_0)}$ will be denoted by $|\cdot|_{s,\Omega}$ and $|\cdot|_{s,\Gamma_0}$, respectively. In particular, the symbols $|\cdot|_{0,\Omega}$ and $|\cdot|_{0,\Gamma_0}$ will represent $\|\cdot\|_{L_2(\Omega)}$ and $\|\cdot\|_{L_2(\Gamma_0)}$, respectively.

2. Stabilization by nonlinear damping on the entire domain

This section is focused on the asymptotic behaviour of solutions to the coupled PDE system (1.1) with dissipation distributed on the entire domain Ω . Regarding the nonlinear function F , we shall assume that

Hypothesis 2.1. *F is a continuous function on the real line such that:*

- (i) *F is monotone strictly increasing, $F(0) = 0$;*
- (ii) *$ms^2 \leq sF(s) \leq Ms^6$ for $|s| \geq 1$, with $0 < m \leq M$.*

We recall that from part (i) of Hypothesis 2.1 it follows that there exists a real valued function $h(x)$ which is defined for $x \geq 0$, it is concave, strictly increasing, with $h(0) = 0$, and it satisfies

$$h(sF(s)) \geq s^2 + F^2(s), \quad \text{for } |s| \leq N, \text{ for some } N > 0. \quad (2.1)$$

Such function can always be constructed, thanks to the monotonicity property assumed on F by Hypothesis 2.1, as explained in [10, p. 510]. With this function one first defines

$$\tilde{h}(x) = h\left(\frac{x}{|Q_{0T}|}\right), \quad x \geq 0, \quad (2.2)$$

where $|Q_{0T}|$ denotes the measure of $Q_{0T} = (0, T) \times Q$. Note that since \tilde{h} is monotone increasing, $cI + \tilde{h}$ is invertible for any constant $c \geq 0$. Then, with K a positive constant, we set

$$p(x) := (cI + \tilde{h})^{-1}(Kx), \quad (2.3)$$

which is readily a continuous, positive, strictly increasing function, with $p(0) = 0$. Finally, let

$$q(x) := x - (I + p)^{-1}(x), \quad x > 0. \quad (2.4)$$

Then, decay rates of the energy $E(t)$ of weak solutions $[\vec{z}(t, \cdot), \vec{v}(t, \cdot)]$ to the PDE model (1.1), as defined by (1.6), are described by the following result.

Theorem 2.2. ([4]) *Assume Hypothesis 2.1. Then the energy $E(t)$ of every weak solution to the coupled system (1.1) decays uniformly to zero, as $t \rightarrow +\infty$. More precisely, there exists a $T_0 > 0$ such that*

$$E(t) \leq s(t/T_0 - 1) \quad \text{for } t > T_0, \quad (2.5)$$

where $\lim_{t \rightarrow +\infty} s(t) = 0$ and $s(t)$ is the solution to the Ordinary Differential Equation

$$\begin{cases} s'(t) + q(s(t)) = 0 \\ s(0) = E(0), \end{cases} \quad (2.6)$$

(where q is as given in (2.4)). Here, the constant K in (2.3) will depend on $E(0)$ and time T_0 , and the constant c (in (2.3)) depends on the measure of Q_{0T_0} .

Remark 2.3. We note that we do not assume *a priori* any kind of growth condition on the nonlinearity F near the origin. Knowledge of the rate growth of the nonlinear function F at the origin allows to obtain more explicit decay rates of the energy (see [10, 9]).

The case of saturated feedback laws. Since the dissipation term $F(z_t)$ in (1.1) can be interpreted as a control in *feedback* form, it is both natural and of significance to include in the present analysis the case of feedback laws F subject to *saturation*. A typical example is given by $F(y) = \min\{1, k/|y|\} y$, with k a positive constant. However, in this case the lower growth condition in Hypothesis 2.1(ii) is not fulfilled and needs to be removed. To accomplish this goal, we aim to study the stability properties of the coupled PDE system (1.1) by simply assuming that

Hypothesis 2.4. *F is a continuous function on the real line such that*

- (i) *F is (monotone) strictly increasing for $|s| \leq 1$, while it is non-decreasing for $|s| \geq 1$; $F(0) = 0$;*
- (ii) *there exists $M > 0$ such that $0 < s F(s) \leq M s^6$ for $|s| \geq 1$.*

Under this weak condition we are able to show that decay rates of the energy $E(t)$ are still uniform, provided that initial data belong to a slightly smoother function space $W_\epsilon \subset Y$, rather than the energy space Y (see [4, Theorem 2.8]).

Remark 2.5. Due to space constraints, the precise statement of the aforementioned stability result will not be included in the present article. It would require the introduction of some preliminary material: first of all, the definition of the new function space W_ϵ and the corresponding new energy $E_1(t)$. In short, Theorem 2.8 in [4] establishes that if the nonlinear function F fulfils the weaker Hypothesis 2.4, then decays rates can be described as well via the solution $s_1(t)$ to a nonlinear ODE (as (2.6) of Theorem 2.2), depending this time on $E_1(0)$ instead of $E(0)$. The new key ingredient in the proof of this result is an estimate of the norm $|z_t|_{\epsilon, \Omega}$ (of solutions to (1.1) with initial data in W_ϵ), which can be shown by using linear and nonlinear interpolation methods (see [4]).

3. Stabilization by a locally distributed damping

In this section we consider the PDE system (1.1) with damping acting only on an arbitrary small layer around the interface Γ_0 . Henceforth, we shall write in this case

$$F(z_t(t, x)) = d(x) g(z_t(t, x)), \quad (3.1)$$

where d is a nonnegative function on Ω which is active only in a neighbourhood of Γ_0 , and g is a real function subject to appropriate conditions (see Hypothesis 3.5 below). Preliminarily, we recall that model (1.1) with *linear* localized damping (i.e. with $g(z_t) = z_t$ in (3.1)) has been studied in [5], in the case when the following geometric assumptions hold true (*cf.* [13]):

Hypothesis 3.1.

- *The domain Ω is convex and the following negatively star shaped condition is satisfied: there exists a point $x_0 \in \mathbb{R}^n$ such that*

$$(x - x_0) \cdot \nu \leq 0 \quad \text{on } \Gamma_1;$$

- *$d(x) \equiv 1$ on $\tilde{\Omega} \subset \Omega$ and $\overline{\tilde{\Omega}} \supset \Gamma_0$.*

Under these conditions it has been shown that the energy $E(t)$ of system (1.1), as defined by (1.6), decays exponentially to zero, as $t \rightarrow +\infty$. More precisely, the following stability result has been established for system (1.1) with $F(z_t) = d(x)z_t$, which we rewrite here for readers' convenience:

$$\left\{ \begin{array}{ll} z_{tt} = \Delta z - d(x)z_t & \text{in } Q \\ \frac{\partial z}{\partial \nu} + d_1 z = 0 & \text{on } \Sigma_1 \\ \frac{\partial z}{\partial \nu} + D_0 z_t + \beta D_0 z = v_t & \text{on } \Sigma_0 \\ z(0, \cdot) = z^0, \quad z_t(0, \cdot) = z^1 & \text{in } \Omega \\ v_{tt} + \mathcal{A}v + \rho \mathcal{A}^\alpha v_t + z_t|_{\Gamma_0} = 0 & \text{on } \Sigma_0 \\ v(0, \cdot) = v^0, \quad v_t(0, \cdot) = v^1 & \text{in } \Gamma_0. \end{array} \right. \quad (3.2)$$

Theorem 3.2. ([5]) *Assume Hypothesis 3.1. Then the energy $E(t)$ of every solution $[\vec{z}, \vec{v}]$ to the coupled system (3.2) decays exponentially to zero, as $t \rightarrow +\infty$, that is*

$$E(t) \leq C e^{-\omega t} E(0), \quad t \geq 0, \quad (3.3)$$

for some positive constants C, ω . The constants C and ω do not depend on $E(0)$, but they depend on $\tilde{\Omega}$. More precisely, $C \rightarrow +\infty$ as $\beta \rightarrow 0$ or the area of the support of d (the "height" of $\tilde{\Omega}$) goes to zero.

Remark 3.3. The proof of Theorem 3.2 is contained in [5]. The exponential decay in (3.3) is achieved by showing the equivalent property

that for T sufficiently large one has $E(T) \leq \xi E(0)$, with $0 < \xi < 1$. In turn, this property follows as a consequence of an appropriate integral estimate of the energy functional, whose proof is rather technical and requires several intermediate steps. The integral estimates leading to the final estimate are obtained by applying the well known multipliers' method. Here a key role is played by suitable multipliers which are constructed by using both appropriate cut-off functions and a non-radial vector field $\mathbf{h} \in [C^2(\overline{\Omega})]^n$ such that

$$\mathbf{h} \cdot \nu \leq 0 \quad \text{on } \Gamma_1, \quad \text{and} \quad J(\mathbf{h}) > 0 \quad \text{in } \overline{\Omega}, \quad (3.4)$$

where $J(\mathbf{h})$ denotes the Jacobian matrix of \mathbf{h} . Existence of \mathbf{h} with the features in (3.4) is guaranteed by Hypothesis 3.1 (*cf.* [13]).

Our present goal is to extend the previous result to a more general model with *nonlinear* localized damping. More precisely, we shall consider the PDE system

$$\begin{cases} z_{tt} = \Delta z - d(x)g(z_t) & \text{in } Q \\ \frac{\partial z}{\partial \nu} + d_1 z = 0 & \text{on } \Sigma_1 \\ \frac{\partial z}{\partial \nu} + D_0 z_t + \beta D_0 z = v_t & \text{on } \Sigma_0 \\ z(0, \cdot) = z^0, \quad z_t(0, \cdot) = z^1 & \text{in } \Omega \\ v_{tt} + \mathcal{A}v + \rho \mathcal{A}^\alpha v_t + z_t|_{\Gamma_0} = 0 & \text{on } \Sigma_0 \\ v(0, \cdot) = v^0, \quad v_t(0, \cdot) = v^1 & \text{in } \Gamma_0, \end{cases} \quad (3.5)$$

where $d(\cdot)$ is the characteristic function of the neighbourhood $\tilde{\Omega}$ of Γ_0 where the damping is active.

Remark 3.4. In the present case where the model displays *locally distributed* damping, we cannot expect to maintain the polynomial growth of the nonlinear term g allowed by Hypothesis 2.1 (up to power five). In fact, as in the case of similar models with *boundary* dissipation (see [10, 2], [9] and references therein) we need to require linear growth at infinity.

Let us assume that

Hypothesis 3.5. *g is a continuous function on the real line such that:*

- (i) *g is monotone strictly increasing, $g(0) = 0$;*
- (ii) *$ms^2 \leq s g(s) \leq Ms^2$ for $|s| \geq 1$, with $0 < m \leq M$.*

Then, the following generalization of Theorem 3.2 holds true.

Theorem 3.6. *Assume Hypotheses 3.1 and 3.5. Then the energy $E(t)$ of every weak solution to the coupled system (3.5) decays uniformly to zero, as $t \rightarrow +\infty$. Decay rates are estimated via the solution $s(t)$ to an appropriate nonlinear ODE, as described by (2.5) and (2.6).*

Proof of Theorem 3.6 (sketch). Here we give an outline of the proof of Theorem 3.6 for which complete details will be available in a separate paper.

A preliminary step: well-posedness. As a preliminary step in the proof of Theorem 3.6, one needs to consider system (3.5) as an abstract evolution equation of the form $y' = Ay$ in an appropriate Hilbert space Y , where A denotes the nonlinear dynamics operator. Then, well-posedness of this system can be shown by applying the theory of *nonlinear semigroups* ([3]) as done in [1, 4]. Moreover, the regularity of solutions corresponding to smooth initial data will contribute to justification of the computations which are performed next.

Basic approach. In order to obtain decay rates estimates of (finite energy) solutions to system (3.5), we shall follow once more the powerful method introduced by the authors of [10] in the study of semilinear wave equations with nonlinear boundary velocity feedbacks and applied subsequently to various linear and nonlinear coupled PDE models (see [9], providing numerous references). Following this approach, our final objective will be to achieve the nonlinear *functional* inequality (3.10) below. To accomplish this goal, a major role will be played by the choice of suitable multipliers, as described in Remark 3.3.

Energy identity. The starting point is to derive the usual energy identity which illustrates the fact that the system is dissipative.

With respect to the PDE system (3.5), the following energy equality holds for all s and T , with $0 \leq s \leq T$:

$$E(T) + 2 \int_0^T (|D_0^{1/2} z_t|_{0,\Gamma_0}^2 + (d(x)g(z_t), z_t)_\Omega + \rho |\mathcal{A}^{\alpha/2} v_t|_{0,\Gamma_0}^2) dt = E(s) \quad (3.6)$$

In particular, $E(T) \leq E(t)$, $\forall t \leq T$.

Estimate of the energy functional. Next, we seek to obtain an integral estimate of the energy functional on a finite time interval $[0, T]$. Initially, estimates are performed separately on each component of the system. The most difficult step is to obtain an estimate of the *wave* energy functional $\int_0^T E_z(t) dt$. It is here where Hypothesis 3.1 is used in a crucial way. Then, by combining the integral inequalities pertaining to the plate and wave energy functionals, and by using Hypothesis 3.5, along with

monotonicity and concavity properties of \tilde{h} as defined by (2.2), one gets the following soughtafter estimate for the coupled system.

With respect to the total energy $E(t)$ of system (3.5) as defined in (1.6), the following inequality holds for all $T > 0$:

$$\begin{aligned} \int_0^T E(t) dt &\leq C_1(E(0) + E(T)) + C_2[c_1 I + \\ &+ |Q_{0T}| \tilde{h}] \left(\int_0^T \left\{ |D_0^{1/2} z_t|_{0,\Gamma_0}^2 + (d(x)g(z_t), z_t)_\Omega + \right. \right. \\ &\quad \left. \left. + \rho |\mathcal{A}^{\alpha/2} v_t|_{0,\Gamma_0}^2 \right\} dt \right) + C_3 lot(z, v), \end{aligned} \quad (3.7)$$

where crucially C_1 does not depend on T , while the expression $lot(z, v)$ includes all terms which are below energy level. More precisely, for some constant C we have that

$$lot(z, v) \leq C \int_0^T \left(|z|_{0,\Omega}^2 + |z|_{\frac{1}{2}-\delta, \Gamma_0}^2 + |v_t|_{-\frac{1}{2}+\delta, \Gamma_0}^2 \right) dt, \quad 0 < \delta < \frac{1}{2}. \quad (3.8)$$

Absorption of lower order terms and final estimate. The lower order terms can be absorbed by means of a (by now) standard nonlinear compactness/uniqueness argument (see [9]). Then, by using the energy identity (and dissipativity property), the following estimate of the energy function is established.

For T large enough, the energy $E(t)$ of every solution to system (3.5) satisfies

$$\begin{aligned} E(T) &\leq C_T(E(0)) [c I + \tilde{h}] \left(\int_0^T \left\{ |D_0^{1/2} z_t|_{0,\Gamma_0}^2 + (d(x)g(z_t), z_t)_\Omega + \right. \right. \\ &\quad \left. \left. + \rho |\mathcal{A}^{\alpha/2} v_t|_{0,\Gamma_0}^2 \right\} dt \right) \end{aligned} \quad (3.9)$$

where the constant $C_T(E(0))$ remains bounded for bounded values of $E(0)$.

Conclusion. By using the energy identity (3.6) in the right hand side of inequality (3.9), we finally attain

$$E(T) + p(E(T)) \leq E(0), \quad (3.10)$$

with the function p as defined by (2.3). We recall, in particular, that p is constructed in terms of \tilde{h} (hence, in terms of h), which depends on the growth rate of g near the origin. Thus, conclusion follows by applying the general result given in [10, Lemma 3.3], here with $s_m = E(mT)$ (see [10, p. 532] and [2, p. 304]).

References

- [1] Avalos, G. and Lasiecka, I. (1996). Differential Riccati equation for the active control of a problem in structural acoustics. *J. Optim. Theory Appl.*, **91**(3):695–728.
- [2] Avalos, G. and Lasiecka, I. (1998). Uniform decay rates for solutions to a structural acoustic model with nonlinear dissipation. *Appl. Math. and Comp. Sci.*, **8**(2):287–312.
- [3] Barbu, V. (1976). *Nonlinear semigroups and differential equations in Banach spaces*. Noordhoff International Publishing, Leyden.
- [4] Bucci, F. Uniform decay rates of solutions to a system of coupled PDEs with nonlinear internal dissipation. *Advances in Differential Equations* (to appear).
- [5] Bucci, F. and Lasiecka, I. (2002). Exponential decay rates for structural acoustic model with overdamping on the interface and boundary layer dissipation. *Applicable Analysis*, **81**(4):977–999.
- [6] Bucci, F., Lasiecka, I., and Triggiani, Roberto. (2002). Singular estimates and uniform stability of coupled systems of hyperbolic/parabolic PDEs. *Abstr. Appl. Anal.*, **7**(4):169–236.
- [7] Cannarsa, P., Gozzi, F., and Soner, H.M. (1993). A dynamic programming approach to nonlinear boundary control problems of parabolic type. *J. Funct. Anal.*, **117**(1):25–61.
- [8] Cannarsa, P. and Tessitore, M.E.. (1996). Infinite-dimensional Hamilton-Jacobi equations and Dirichlet boundary control problems of parabolic type. *SIAM J. Control Optim.*, **34**(6):1831–1847.
- [9] Lasiecka, I. (2002). *Mathematical Control Theory of Coupled PDEs*, CBMS-NSF Regional Conference Series in Applied Mathematics, **75**, SIAM, Philadelphia.
- [10] Lasiecka, I. and Tataru, D. (1993). Uniform boundary stabilization of semilinear wave equations with nonlinear boundary damping. *Differential and Integral Equations*, **6**(3):507–533.
- [11] Lasiecka, I. and Triggiani, R. (2000). *Control Theory for Partial Differential Equations: Continuous and Approximation Theories, Vol. I: Abstract Parabolic Systems; Vol. II: Abstract Hyperbolic-like Systems over a Finite Time Horizon*. Encyclopedia of Mathematics and its Applications, Vol. **74-75**, Cambridge University Press.
- [12] Lasiecka, I. and Triggiani, R. (2001). Optimal control and Algebraic Riccati Equations under singular estimates for $e^{At}B$ in the absence of analiticity. Part I: The stable case. In *Differential Equations and Control Theory*, Marcel Dekker Lecture Notes in Pure and Applied Mathematics, **225**:193–219.
- [13] Lasiecka, I., Triggiani, R., and Zhang, X. (2000). Nonconservative wave equations with unobserved Neumann boundary conditions: global uniqueness and observability in one shot. *Contemporary Mathematics*, **268**:227–325.
- [14] Martinez, P. and Vancostenoble, J. (2000). Exponential stability for the wave equation with weak nonmonotone damping, *Port. Math.*, **57**(3):285–310.

EXISTENCE OF STRONG SOLUTIONS OF FULLY NONLINEAR ELLIPTIC EQUATIONS

Adriana Buică

Department of Applied Mathematics

Babeş-Bolyai University of Cluj-Napoca,

1 Kogalniceanu str., RO-3400 Romania

abuica@math.ubbcluj.ro

Abstract The aim of this paper is to study the solvability of the Dirichlet problem for certain types of fully nonlinear elliptic equations. The theory of weakly-near operators, combined to Contraction Mapping and Schauder fixed point theorems, is used. Our main results generalizes similar ones given by S. Campanato and A. Tarsia.

Keywords: nonlinear, elliptic, Sobolev spaces, weakly-near operators, fixed point.

1. Introduction

In this paper we study the solvability of the Dirichlet problem for certain types of fully nonlinear elliptic equations of the form

$$u \in H^2(\Omega) \cap H_0^1(\Omega), \quad a(x, u, Du, D^2u) = f(x), \quad \text{for a.e. } x \in \Omega. \quad (1.1)$$

In what follows, Ω will be a C^2 bounded domain of \mathbb{R}^n . We denote by \mathcal{M}_n the space of $n \times n$ real matrices; $|\cdot|_m$ is the euclidean norm in \mathbb{R}^m ; $trN = \sum_{i=1}^n \xi_{ii}$ is the trace of the $n \times n$ matrix $N = (\xi_{ij})$. The Sobolev spaces $H^2(\Omega)$ and $H_0^1(\Omega)$ are as defined in Adams [1975].

We assume that the function $a : \Omega \times \mathbb{R} \times \mathbb{R}^n \times \mathcal{M}_n \rightarrow \mathbb{R}$ fulfills the following conditions:

- (a1) $a(x, 0, 0, 0) = 0$,
- (a2) $a(\cdot, r, d, M)$ is measurable,
- (a3) $a(x, \cdot, \cdot, \cdot)$ is continuous,
- (a4) there exist $\alpha, \beta, \gamma \geq 0$ such that $|a(x, r, d, M)| \leq \alpha|r| + \beta|d|_n + \gamma|M|_n^2$,
for all $r \in \mathbb{R}$, $d \in \mathbb{R}^n$, $M \in \mathcal{M}_n$ and for a.e. $x \in \Omega$.

The following ellipticity condition is satisfied.

(a5) there exist $c_1, c_2, c_3 > 0$ with $0 < c := c_1 - c_2 - c_3 < 1$ such that $[a(x, r, d, N+M) - a(x, r, d, M)]trN \geq c_1|trN|^2 - c_2|trN| \cdot |N|_{n^2} - c_3|N|_{n^2}^2$, for all $r \in \mathbb{R}$, $d \in \mathbb{R}^n$, $M, N \in \mathcal{M}_n$ and for a.e. $x \in \Omega$.

We obtain an existence and uniqueness result and another existence result. The theory of weakly-near operators (see Buică and Domokos [2002]), combined to Contraction Mapping Theorem or Schauder Fixed Point Theorem, is used. Our main results are the following.

Theorem 1.1 *Let us assume that the function $a : \Omega \times \mathbb{R} \times \mathbb{R}^n \times \mathcal{M}_n \rightarrow \mathbb{R}$ satisfies (a1)-(a5), and there exist $l_1, l_2 > 0$ such that*

$$|a(x, r, d, M) - a(x, s, \delta, M)| \leq l_1|r - s| + l_2|d - \delta|, \quad (1.2)$$

for a.e. $x \in \Omega$ and for all $r, s \in \mathbb{R}$, $d, \delta \in \mathbb{R}^n$, $M \in \mathcal{M}_n$. If $\frac{l_1}{\lambda_1} + \frac{l_2}{\sqrt{\lambda_1}} < c$ then (1.1) has a unique solution.

Theorem 1.2 *Let us assume that the function $a : \Omega \times \mathbb{R} \times \mathbb{R}^n \times \mathcal{M}_n \rightarrow \mathbb{R}$ satisfies (a1)-(a5), and there exist $l_1, l_2 > 0$ such that*

$$|a(x, r, d, M) - a(x, 0, 0, M)| \leq l_1|r| + l_2|d|, \quad (1.3)$$

for a.e. $x \in \Omega$ and for all $r \in \mathbb{R}$, $d \in \mathbb{R}^n$, $M \in \mathcal{M}_n$. If $\frac{l_1}{\lambda_1} + \frac{l_2}{\sqrt{\lambda_1}} < c$ then (1.1) has at least one solution.

We extend and generalize similar results of S. Campanato [1989] and A. Tarsia [1996, 1998]. Another related results are given by D.K. Palagachev [1993], R. Precup [1995, 1997], A. Buică and A. Domokos [2001], A. Buică and F. Aldea [2000], A. Buică [2001b].

The next sections are: 2. Theoretical preliminaries (necessary results from the theory of weakly-near operators and elliptic equations are presented), 3. Proof of main results and 4. Comments (we explain the relations between our results and those existing in the literature).

2. Theoretical preliminaries

Let X be a nonempty set and Z be a Banach space. Let $A, B : X \rightarrow Z$ be two operators. S. Campanato introduced the following notion of nearness between operators in order to use it in the study of fully nonlinear elliptic equations (see Campanato [1989, 1993], Tarsia [1996, 1998]).

Definition 2.1 We say that A is near B if there exists $\alpha > 0$ and $0 \leq c < 1$ such that

$$\|Bx - By - \alpha(Ax - Ay)\| \leq c\|Bx - By\|, \text{ for all } x, y \in X. \quad (2.1)$$

In a joint paper with A. Domokos, we generalized this notion using an accretivity-type condition, instead of a contraction-type one.

Let us denote by Φ the set of all functions $\varphi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$, such that $\varphi(0) = 0$, $\varphi(r) > 0$ for $r > 0$, $\liminf_{r \rightarrow \infty} \varphi(r) > 0$ and $\liminf_{r \rightarrow r_0} \varphi(r) = 0$ implies $r_0 = 0$. In this paper we shall refer only to the functions φ in Φ .

We say that A is φ -accretive with respect to B , if for every $x, y \in X$ there exists $j(Bx - By) \in J(Bx - By)$ such that

$$\langle Ax - Ay, j(Bx - By) \rangle \geq \varphi(\|Bx - By\|) \|Bx - By\|, \quad (2.2)$$

where $J : Z \rightsquigarrow Z^*$ is the normalized duality map of Z .

The map A is continuous with respect to B if $A \circ B^{-1} : B(X) \rightsquigarrow Z$ has a continuous selection.

The next definition introduce the weak-nearness notion.

Definition 2.2 We say that A is weakly-near B if A is φ -accretive with respect to B and continuous with respect to B .

This notion extends the property of the differential operator to be "near" (or to "approximate") the map, as well as other approximation notions used in nonsmooth theory of inverse or implicit functions (for details in this direction, see Buică and Domokos [2002], Domokos [1997, 2000], Buică [2001b]). The next results will be used in Section 3. They are taken from Buică [2001], Buică and Domokos [2002].

Proposition 2.1 Let A be weakly-near to B . If B is bijective, then A is bijective.

Let $z \in Z$ and $A_1, A_2 : X \rightarrow Z$ be two mappings. Let us consider the equation $A_1(x) = z$, whose solvability is assured by the weak-nearness between the operator A_1 and a bijective operator $B : X \rightarrow Z$. Let x_1^* be a solution of this equation. Let us consider, also, the equation $A_2(x) = z$, which is assumed to be solvable. Let x_2^* be a solution. In the following theorem we shall give an estimation for "the distance" between x_1^* and x_2^* . This distance depends on the operator B .

Theorem 2.1 Let us assume that the following conditions are fullfilled.

- (i) B is bijective;
- (ii) A_1 is weakly-near to B with $\varphi(t) = \alpha t$, $0 < \alpha < 1$;
- (iii) equation $A_2(x) = z$ has at least one solution.

Then we have the estimation $\|B(x_1^*) - B(x_2^*)\| \leq \frac{1}{\alpha} \|A_1(x_2^*) - A_2(x_2^*)\|$. If, in addition, there exists $\eta > 0$ such that $\|A_1(x) - A_2(x)\| \leq \eta$ for all $x \in X$, then $\|B(x_1^*) - B(x_2^*)\| \leq \frac{1}{\alpha} \eta$.

We also need the following lemmas which are taken from Gilbarg and Trüdinger [1983], Precup [1997].

Lemma 2.1 *The Laplace operator $\Delta : H^2(\Omega) \cap H_0^1(\Omega) \rightarrow L^2(\Omega)$ is well defined and it is a homeomorphism.*

Lemma 2.2 *For every $u \in H^2(\Omega) \cap H_0^1(\Omega)$,*

$$\|u\|_{L^2} \leq \frac{1}{\lambda_1} \|\Delta u\|_{L^2}, \quad \|Du\|_{L^2} \leq \frac{1}{\sqrt{\lambda_1}} \|\Delta u\|_{L^2}.$$

Lemma 2.3 *Let $u \in H^2(\Omega) \cap H_0^1(\Omega)$. Then $\|D^2u\|_2 = \|\Delta u\|_2$.*

3. Proof of main results

Proof of Theorem 1.1. Let $w \in H^1(\Omega)$. Let us consider the mapping A_w defined by

$$A_w : H^2(\Omega) \cap H_0^1(\Omega) \rightarrow L^2(\Omega), \quad A_w(u)(x) = a(x, w, Dw, D^2u).$$

Let us consider, also, the equation

$$A_w(u) = f. \quad (3.1)$$

1) A_w is well defined and continuous.

Using condition (a4) we obtain that

$$|a(x, u, Du, D^2u)| \leq \alpha|u(x)| + \beta|Du(x)|_n + \gamma|D^2u|_{n^2},$$

for every $u \in H^2(\Omega) \cap H_0^1(\Omega)$.

Because the right hand side of this inequality is an L^p -function, we can deduce that A is well-defined and continuous.

2) A_w is weakly near to Δ .

The mapping A_w is continuous with respect to $B = \Delta$ because A_w and B^{-1} are continuous. We shall prove that A_w is strongly accretive (in fact, strongly monotone) with respect to B . Using (a5) and Lemma 2.3 we obtain:

$$\begin{aligned} & \langle A_w u - A_w v, \Delta u - \Delta v \rangle_{L^2} = \\ &= \int_{\Omega} [a(x, w, D^2u) - a(x, w, D^2v)] \cdot \Delta(u - v) dx \geq \\ &\geq \int_{\Omega} c_1 |\Delta(u - v)|^2 - c_2 |\Delta(u - v)| \cdot |D^2(u - v)| - c_3 |D^2(u - v)|^2 dx \geq \\ &\geq (c_1 - c_2 - c_3) \|\Delta(u - v)\|_{L^2}^2. \end{aligned}$$

Thus

$$\langle A_w(u) - A_w(v), \Delta u - \Delta v \rangle_{L^2} \geq c \|\Delta u - \Delta v\|_{L^2}^2. \quad (3.2)$$

A_w is weakly near to B , which is a bijective map. Then, using Proposition 2.1, A_w is bijective. Thus, equation (3.1) has a unique solution, let us denote it by u_w .

Let us consider now another operator, related to equation (3.1), defined by

$$\mathcal{U} : H^1(\Omega) \rightarrow L^2(\Omega), \quad \mathcal{U}(w) = -\Delta u_w.$$

Let us notice that $v = -\Delta w$ is a fixed point of

$$\mathcal{U} \circ (-\Delta)^{-1} : L^2(\Omega) \rightarrow L^2(\Omega)$$

if and only if $u_w = w$, which means that u_w is a solution of the problem (1.1).

Let $w_1, w_2 \in H^1(\Omega)$ and let us consider the mappings A_{w_1} and A_{w_2} . For every $u \in H^2(\Omega) \cap H_0^1(\Omega)$, using relation (1.2) and Lemma 2.2, we obtain the estimations,

$$\begin{aligned} & \|A_{w_1}(u) - A_{w_2}(u)\|_{L^2} = \\ &= \|a(\cdot, w_1(\cdot), Dw_1(\cdot), D^2u(\cdot)) - a(\cdot, w_2(\cdot), Dw_2(\cdot), D^2u(\cdot))\|_{L^2} \leq \\ &\leq \|l_1|w_1(\cdot) - w_2(\cdot)| + l_2|Dw_1(\cdot) - Dw_2(\cdot)|\|_{L^2} \leq \\ &\leq l_1\|w_1 - w_2\|_{L^2} + l_2\|Dw_1 - Dw_2\|_{L^2} \leq \\ &\leq \left(\frac{l_1}{\lambda_1} + \frac{l_2}{\sqrt{\lambda_1}} \right) \|\Delta w_1 - \Delta w_2\|_{L^2} := \eta. \end{aligned}$$

We apply Theorem 2.1 and obtain $\|B(u_{w_1}) - B(u_{w_2})\|_{L^2} \leq \frac{1}{c}\eta$, which means that

$$\|\mathcal{U}(w_1) - \mathcal{U}(w_2)\|_{L^2} \leq \frac{1}{c} \left(\frac{l_1}{\lambda_1} + \frac{l_2}{\sqrt{\lambda_1}} \right) \cdot \|\Delta w_1 - \Delta w_2\|_2.$$

Then $\mathcal{U} \circ (-\Delta)^{-1}$ is a contraction on $L^2(\Omega)$. By Contraction Mapping Theorem, it has a unique fixed point, v^* . If we denote by $w^* = (-\Delta)^{-1}v^*$, then u_{w^*} is the unique solution of (1.1). ■

Proof of Theorem 1.2. It is possible to consider again the mapping \mathcal{U} like in the proof of Theorem 1.1. Let us notice that w is a fixed point of

$$(-\Delta)^{-1} \circ \mathcal{U} : H^1(\Omega) \rightarrow H^1(\Omega)$$

if and only if u_w is a solution of the problem (1.1). This time we shall apply the Schauder Fixed Point Theorem.

First we shall prove that \mathcal{U} is continuous in every $v \in H^1(\Omega)$. Let us denote by $u_v \in H^2(\Omega) \cap H_0^1(\Omega)$ the unique solution of $A_v(u) = f$, with A_v defined like in the proof of previous theorem. The hypotheses (a3) and (a4) assure that the mapping $w \mapsto a(\cdot, w, Dw, D^2u_v)$ is continuous from $H^1(\Omega)$ to $L^2(\Omega)$. In particular, is continuous in v . Then, for every $\varepsilon > 0$ there exists $\delta > 0$ such that, whenever $w \in H^1(\Omega)$, $\|w - v\|_{H^1} < \delta$,

$$\|a(\cdot, w, Dw, D^2u_v) - a(\cdot, v, Dv, D^2u_v)\|_{L^2} < \varepsilon.$$

Then, $\|A_w(u_v) - A_v(u_v)\|_{L^2} \leq \varepsilon$. We apply Theorem 2.1 like in the proof of the previous theorem and obtain $\|\mathcal{U}(w) - \mathcal{U}(v)\|_{L^2} \leq \frac{1}{c} \cdot \varepsilon$. Hence, \mathcal{U} is continuous, indeed.

Let us consider now the following norm in $H^1(\Omega)$,

$$\|w\|_* = l_1 \|w\|_{L^2} + l_2 \|Dw\|_{L^2},$$

which is equivalent to the usual norm,

$$\|w\|_{H^1} = \left(\int_{\Omega} |w|^2 + |Dw|^2 dx \right)^{1/2}.$$

Let $w \in H^1(\Omega)$ and let us consider the mappings A_w and A_0 . For every $u \in H^2(\Omega) \cap H_0^1(\Omega)$ we obtain the following estimations like in the proof of the previous theorem,

$$\|A_w(u) - A_0(u)\|_{L^2} \leq l_1 \|w\|_{L^2} + l_2 \|Dw\|_{L^2} = \|w\|_*.$$

We apply Theorem 2.1 and obtain $\|\mathcal{U}(w) - \mathcal{U}(0)\|_{L^2} \leq \frac{1}{c} \|w\|_*$. Then

$$\|\mathcal{U}(w)\|_{L^2} \leq \frac{1}{c} \|w\|_* + \|\mathcal{U}(0)\|_{L^2}. \quad (3.3)$$

This assures that $\mathcal{U} : (H^1(\Omega), \|\cdot\|_*) \rightarrow (L^2(\Omega), \|\cdot\|_{L^2})$ is a bounded operator. Also, we have that $(-\Delta)^{-1} : L^2(\Omega) \rightarrow H^2(\Omega) \cap H_0^1(\Omega)$ is bounded and $H^2(\Omega)$ is compactly imbedded in $H^1(\Omega)$. Hence, $(-\Delta)^{-1} \circ \mathcal{U} : H^1(\Omega) \rightarrow H^1(\Omega)$ is completely continuous.

Now we prove that there exists an invariant set of $(-\Delta)^{-1} \circ \mathcal{U}$. The following relations hold for every $w \in H^1(\Omega)$,

$$\|(-\Delta)^{-1} \circ \mathcal{U}(w)\|_* \leq \left(\frac{l_1}{\lambda_1} + \frac{l_2}{\sqrt{\lambda_1}} \right) \|\mathcal{U}(w)\|_{L^2} \leq l_3 \|w\|_* + c l_3 \|\mathcal{U}(0)\|_{L^2},$$

where $l_3 := \frac{1}{c} \left(\frac{l_1}{\lambda_1} + \frac{l_2}{\sqrt{\lambda_1}} \right)$. We have used the definitions of \mathcal{U} and $\|w\|_*$, Lemma 2.2 and relation (3.3). Since $l_3 < 1$, we let

$$R \geq \frac{c l_3 \|\mathcal{U}(0)\|_{L^2}}{1 - l_3}.$$

We have that $\|(-\Delta)^{-1} \circ \mathcal{U}(w)\|_* \leq R$ whenever $\|w\|_* \leq R$, i.e. the ball centered in origin with radius R from the Banach space $(H^1(\Omega), \|\cdot\|_*)$, is an invariant set for the mapping $(-\Delta)^{-1} \circ \mathcal{U}$.

Applying the Schauder fixed point theorem we deduce that $(-\Delta)^{-1} \circ \mathcal{U}$ has at last one fixed point, w^* . Then u_{w^*} is a solution of (1.1). ■

4. Comments

In this section we relate our results to some other ones in the literature. S. Campanato [1989] and A. Tarsia [1996, 1998] considered a function $a = a(x, M)$ (i.e. the equation (1.1) does not depend explicitly on the function u and its gradient) which satisfies (a1)-(a3) and the following ellipticity condition:

(a6) there exist three positive constants α, β, γ , with $\gamma + \delta < 1$ such that

$$|trN - \alpha[a(x, M + N) - a(x, M)]| \leq \gamma|N|_{n^2} + \delta|trN|,$$

for almost every $x \in \Omega$, for all $M, N \in \mathcal{M}_n$.

This is stronger than (a5)+(a4). Hence, the particular case of Theorem 1.1 when $a = a(x, M)$ generalizes the main result in Campanato [1989].

In Tarsia [1998] the following problem is also considered

$$u \in H^2(\Omega) \cap H_0^1(\Omega), \quad a(x, D^2u) + g(x, u) = f(x), \quad \text{for a.e. } x \in \Omega. \quad (4.4)$$

The function $g : \Omega \times \mathbb{R} \rightarrow \mathbb{R}$ is measurable and there exists $l \leq \lambda_1$ (where λ_1 is the smallest positive eigenvalue of the operator $-\Delta$), and for all $r, s \in \mathbb{R}$ and for a.e. $x \in \Omega$

- (i) $g(x, 0) = 0$,
- (ii) $0 \leq [g(x, r) - g(x, s)](r - s)$,
- (iii) $|g(x, r) - g(x, s)| \leq l|r - s|$.

In Tarsia [1998] an existence and uniqueness result is obtained for (4.4) when l is sufficiently small. It is easily seen that this result is also a consequence of Theorem 1.1.

Campanato [1989] and Tarsia [1998] used the theory of near-operators. The theory of weakly-near operators permitted us to deal with the weaker ellipticity condition (a5). D. Palagachev [1993] studied the solvability of the Dirichlet problem for a class of fully nonlinear elliptic equations under ellipticity condition (a6). The equation contained a term of the form $f = f(x, u, Du)$. The author combined the theory of near-operators to the Leray-Schauder Fixed Point Theorem. We gave some existence and uniqueness results in $W^{2,p}(\Omega) \cap W_0^1(\Omega)$ (with an arbitrary $p > 1$) for equation $a(x, u, Du, D^2u) = f$ (see Buică and Domokos [2002]). We used another ellipticity condition which assured the weak nearness to a general linear elliptic operator. The results of this paper extends similar ones appeared in Buică [2001b], where the gradient did not appear in the form of the equation. Our joint paper, Buică and Aldea [2000], contains a data dependence theorem for equations of the form $a(x, D^2u) = f$ in $H^2(\Omega) \cap H_0^1(\Omega)$. Also these results extends to the case of fully nonlinear equations those given by R. Precup [1995, 1997] for semilinear elliptic equations.

References

- [1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York/San Francisco/London, 1975.
- [2] A. Buică and F. Aldea, On Peetre's condition in the coincidence theory, *Proc. of Séminaire de la théorie de la meilleure approximation*, Cluj-Napoca, pg. 17-27, october 2000.
- [3] A. Buică, Some properties preserved by weak nearness, *Seminar on fixed point theory Cluj-Napoca* 2: 65-71, 2001.
- [4] A. Buică, *Principii de coincidență și aplicații* (Coincidence Principles and Applications, in romanian), Cluj University Press, Cluj-Napoca, 2001.
- [5] A. Buică and A. Domokos, Nearness, accretivity and the solvability of nonlinear equations, *Numer. Funct. Anal. Appl.* 23:477-493, 2002.
- [6] S. Campanato, A Cordes type condition for nonlinear variational systems, *Rend. Acc. Naz. delle Sc.* 107: 307-321, 1989.
- [7] S. Campanato, Further contribution to the theory of near mappings, *Le Matematiche* 48: 183-187, 1993.
- [8] A. Domokos, *Teoreme de funcții implicate nenedede și aplicațiile lor* (Nonsmooth implicit functions and their applications, in romanian), Babeș-Bolyai University, Cluj-Napoca, Doctoral thesis, 1997.
- [9] A. Domokos, Implicit function theorems for m-accretive and locally accretive set-valued mappings, *Nonlinear Analysis* 41: 221-241, 2000.
- [10] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer Verlag, 1983.
- [11] D.K. Palagachev, Global strong solvability of Dirichlet problem for a class of nonlinear elliptic equations in the plane, *Le Matematiche* 48: 311-321, 1993.
- [12] R. Precup, Existence results for nonlinear boundary value problems under non-resonance conditions, *Qualitative Problems for Differential Equations and Control Theory* (C. Corduneanu, ed.) World Scientific, Singapore, pp. 263-273, 1995.
- [13] R. Precup, *Ecuatii cu Derivate Partiale*, Transilvania Press, Cluj, 1997.
- [14] A. Tarsia, Some topological properties preserved by nearness between operators and applications to PDE, *Czech. Math. J.* 46(121): 607-624, 1996.
- [15] A. Tarsia, Differential equations and implicit function: a generalization of the near operators theorem, *Topol. Meth. Nonlin. Anal.* 11: 115-133, 1998.

FREE BOUNDARY CONDITIONS FOR INTRINSIC SHELL MODELS

John Cagnol

*Pôle Universitaire Léonard de Vinci,
Paris, France*
John.Cagnol@devinci.fr

Catherine Lebiedzik*

*University of Virginia and Pole Universitaire Léonard de Vinci,
Charlottesville, Virginia, USA*
cgl5f@virginia.edu

Abstract We derive the free boundary conditions and associated strong form of a shallow Kirchhoff shell model based on the intrinsic geometry methods of Michael Delfour and Jean-Paul Zolésio. Manipulations with the model result in a cleaner form where the displacement of the shell and shell boundary is written explicitly in terms of standard tangential operators.

1. Introduction

A wide variety of engineering and applied mathematics problems can be modeled in the context of thin shells. Many problems involve issues such as, for example, boundary feedback control, for which sophisticated mathematical analyses have been developed in the analogous plate case. The difficulty with adapting these techniques to classical shell models is one motivating factor driving the development of new modeling schemes.

The technique proposed by Michel Delfour and Jean-Paul Zolésio [6], [7], and used here, takes advantage of the intrinsic geometric properties of the shell. The model used here was introduced in [5] where the Kirchhoff hypothesis and shallowness assumption were used. We here continue our previous work on the model by deriving the free boun-

*Research supported by the National Science Foundation under Grant INT-0104431.

dary conditions and associated strong from of a Kirchoff-type model presented in the language of the intrinsic geometry of Delfour and Zolésio [8], [6]. The free boundary conditions are very important from both a physical and a control point of view. These are the natural (Neumann-type) boundary conditions and correspond to the actions of bending moments and shear forces. One can apply forces and torques through these moments in order to control and stabilize the vibrating body. This has been studied extensively in the case of plates[13], [14].

Classical shell modeling is a subject well covered in many books, so here we will mention only results which are related to the work at hand. The method of integrating over thickness to derive an equation for the mid-surface of a plate or shell was pioneered by Kirchhoff [10] and Love [16]; and developed and improved by Koiter[11], [12]. The important works of Bernadou, Boissière, Ciarlet, Miara [1], [2], [3], [4] and others analyzed Koiter's model in detail and established ellipticity of the strain energy, convergence and error estimates allowing application of conforming and non-conforming finite element methods. However, the appearance of variable coefficients in the principal part of the operator (a necessary feature of shell models) often makes theoretical analysis difficult, especially from the point of view of stabilization and control. Thus, shell models based on more geometrical arguments are currently being proposed to circumvent these problems.

The technique proposed by Michel Delfour and Jean-Paul Zolésio [6], [7], [8] (and used here) takes advantage of the intrinsic geometric properties of the shell. Here, the shell is described in terms of tangential differential operators which are defined by means of the oriented boundary distance function in \mathbb{R}^3 . Sobolev spaces, Green's formula, and key inequalities such as Poincaré and Korn's inequalities are all well-defined. Delfour and Zolésio have constructed models under a variety of assumptions. The model developed in this paper was introduced in [5] where the Kirchhoff hypothesis and shallowness criterion were used.

This paper expands the shell model introduced in [5] by formulating the associated 'free' boundary conditions. Though this model was used quite successfully for an application involving uniform stability of a structural acoustic model, clearly we assumed a fully clamped shell boundary in the derivation. In contrast, in this work we assume that the shell is only clamped on a *portion* of the boundary, and free on the rest. The assumption that the shell is partially clamped is necessary to conclude that the elastic energy is still, in fact, coercive. In order to derive the boundary conditions we explicitly calculate the integrals over the boundary using the appropriate form of Green's formula. The

final result is presented in terms of standard tangential operators, with coupling and lower-order terms separated in a condensed way.

2. Intrinsic Geometric Modeling

We here include a brief discussion of the oriented distance function and the intrinsic geometric methods of Delfour and Zolésio. The reader is referred to [8], [6] for a definitive exposition on this topic. Consider a domain $\mathcal{O} \subset \mathbb{R}^3$ whose nonempty boundary $\partial\mathcal{O}$ is a C^3 two-dimensional submanifold of \mathbb{R}^3 . Define the oriented (or signed) distance function to \mathcal{O} as

$$b(x) = d_{\mathcal{O}}(x) - d_{\mathbb{R}^3 \setminus \mathcal{O}}(x) \quad (1)$$

where d is the Euclidean distance from the point x to the domain \mathcal{O} . In other words, $b(x)$ is simply the positive or negative distance to the boundary $\partial\mathcal{O}$, depending on if we are outside or inside the domain \mathcal{O} . It can be shown that for every $x \in \partial\mathcal{O}$, there exists a neighborhood where the function $\nabla b = \nu$, the unit outward external normal to $\partial\mathcal{O}$ [6]. Consider a subset $\Gamma \subseteq \partial\mathcal{O}$ which will eventually become the mid-surface of our shell. We define the projection $p(x)$ of a point x onto Γ as $p(x) = x - b(x)\nabla b(x)$. Then, we define a shell S_h of thickness h as

$$S_h(\Gamma) \equiv \{x \in \mathbb{R}^3 : p(x) \in \Gamma, |b(x)| < h/2\} \quad (2)$$

Let $\Upsilon \equiv \partial\Gamma$ denote the boundary of Γ . and (X, z) be the natural curvilinear coordinate system induced on the shell S_h , where the coordinate vector X gives the position of a point on the mid-surface Γ , and $z \in (-\frac{h}{2}, \frac{h}{2})$ gives the vertical (normal) distance from the mid-surface. Using this notation, we also define the “flow mapping” $T_z(X)$ as $T_z(X) = X + z\nabla b(X)$ for all X and z in S_h . This allows us to reconstruct the action at a given height z of the shell, once we know the action of the midsurface Γ . Define as Γ^z the surface $T_z(\Gamma)$ at the ‘altitude’ z . Then, one can also describe the shell S_h as $S_h = \bigcup_{z=-h/2}^{h/2} \Gamma^z$. The curvatures of the shell will be denoted H and K . These can be reconstructed from the boundary distance function $b(x)$ by noting that at any point (X, z) , the matrix D^2b has eigenvalues $0, \lambda_1, \lambda_2$. The curvatures are then given by $\text{tr}(D^2b) = 2H = \lambda_1 + \lambda_2$ and $K = \lambda_1\lambda_2$.

Next, we mention briefly some useful aspects of the tangential differential calculus. Given $f \in C^1(\Gamma)$, we define the tangential gradient ∇_Γ of the scalar function f by means of the projection as

$$\nabla_\Gamma f \equiv [\nabla(f \circ p)(x)]_\Gamma \quad (3)$$

This notion of the tangential gradient is equivalent to the classical definition using an extension F of f in the neighborhood of Γ , i.e.

$\nabla_\Gamma f = \nabla F|_\Gamma - \frac{\partial F}{\partial \nu} \nu$ [6]. Following the same idea we can define the tangential Jacobian matrix of a vector function $v \in C^1(\Gamma)^3$ as

$$D_\Gamma v \equiv D(v \circ p)|_\Gamma \text{ or } (D_\Gamma v)_{ij} = (\nabla_\Gamma v_i)_j$$

the tangential divergence as $\operatorname{div}_\Gamma v \equiv \operatorname{div}(v \circ p)|_\Gamma$, the Hessian $D_\Gamma^2 f$ of $f \in C^2(\Gamma)$ as $D_\Gamma^2 f = D_\Gamma(\nabla_\Gamma f)$, the Laplace-Beltrami operator of $f \in C^2(\Gamma)$ as $\Delta_\Gamma f \equiv \operatorname{div}_\Gamma(\nabla_\Gamma f) = \Delta(f \circ p)|_\Gamma$, the tangential linear strain tensor of elasticity as $\varepsilon_\Gamma(v) \equiv \frac{1}{2}(D_\Gamma v + {}^*D_\Gamma v) = \varepsilon(v \circ p)|_\Gamma$, and the tangential vectorial divergence of a second-order tensor A as $\operatorname{div}_\Gamma A \equiv \operatorname{div}(A \circ p)|_\Gamma = \operatorname{div}_\Gamma A_i$. Using the definitions given in section (2) one can derive Green's formula in the tangential calculus [6]:

$$\int_\Gamma f \operatorname{div}_\Gamma v \, d\Gamma + \int_\Gamma \langle \nabla_\Gamma f, v \rangle \, d\Gamma = \int_\Upsilon \langle fv, \nu \rangle \, d\Upsilon + 2 \int_\Gamma f H \langle v, \nabla b \rangle \, d\Gamma \quad (4)$$

where ν is the outward unit normal to the curve Υ . Finally, from [6, 8] we have that

$$\langle \nabla_\Gamma w, \nabla b \rangle = 0, \quad D_\Gamma v \nabla b = 0 \quad (5)$$

by definition for any scalar w and vector v . In addition, if we consider a purely tangent vector $v = v_\Gamma$, i.e. $\langle v_\Gamma, \nabla b \rangle = 0$, we can take the tangential gradient of both sides of this expression and derive

$$D^2 b v_\Gamma + {}^*D_\Gamma v_\Gamma \nabla b = 0 \quad (6)$$

In this case the first term of equation (4) is also zero. We shall note $|w|_{s,\Gamma} \equiv |w|_{H^s(\Gamma)}$ and $(u, v)_\Gamma \equiv \int_\Gamma uv \, d\Gamma$ where $\langle \cdot, \cdot \rangle$ denotes the scalar product of two vectors. Throughout this paper the conventions of [9] concerning tensors are used.

Hypothesis 2.1 *The following assumptions are imposed on the shell S_h with midsurface Γ .*

- (i) *The shell is assumed to be made of an isotropic and homogeneous material, so that the Lamé coefficients $\lambda > 0$ and $\mu > 0$ are constant.*
- (ii) *The thickness h of the shell is small enough to accommodate the curvatures H and K , i.e. the product of the thickness by the curvatures is small as compared to 1. As a consequence we shall drop terms of order equal or greater than 2 in the series expansions with respect to the radial variable. This also implies that $j(z) = \det(DT_z) = \det(I - zD^2 b) = 1$.*

- (iii) (Kirchhoff Hypothesis) Let T be a transformation of the shell S_h , and let $\mathbf{e} = (e_\Gamma, w)$ be the corresponding transformation of the mid-surface. In the classical thin plate theory named after Kirchhoff, the displacement vectors T and $e \circ p$ are related by the hypothesis that the filaments of the plate initially perpendicular to the middle surface remain straight and perpendicular to the deformed surface, and undergo neither contraction nor extension. We may generalize this hypothesis to the case of a shell using the intrinsic geometry, yielding $T = \mathbf{e} \circ p - b (*D_{\Gamma_0} \mathbf{e} \nabla b) \circ p$
- (iv) We will assume the boundary Υ consists of two open connected regions Υ_0 and Υ_1 , with $\Upsilon = \overline{\Upsilon_0 \cup \Upsilon_1}$ and $\emptyset = \Upsilon_0 \cap \Upsilon_1$. We will clamp the shell on Υ_0 , and allow Υ_1 to be free. Υ_1 may be empty, in which case we recover the result of [5].

We denote by \mathbf{e} the transformation of the shell mid-surface and by e_Γ and e_n the tangential and normal components of \mathbf{e} in local coordinates. We define w to be the magnitude of the normal displacement.

$$w = \langle \mathbf{e}, \nabla b \rangle, \quad e_n = w \nabla b, \quad e_\Gamma = \mathbf{e} - e_n \quad (7)$$

Here we list the following definitions and properties derived in [5]:

Lemma 1 *The following strain-displacement relation holds for a shell modeled in the intrinsic geometry under Hypothesis 2.1 (i)-(iii).*

$$\begin{aligned} \varepsilon(T) &= (\varepsilon_\Gamma(e_\Gamma) + w D^2 b + V_\Gamma e_\Gamma) \circ p \\ &\quad - b (-\varepsilon_\Gamma(D^2 b e_\Gamma) + C_\Gamma e_\Gamma + S_\Gamma w + G_\Gamma w + w(D^2 b)^2) \circ p \end{aligned} \quad (8)$$

where ε_Γ is the tangential linear strain tensor of elasticity and C_Γ , V_Γ , G_Γ , and S_Γ are defined by

$$C_\Gamma u = \frac{1}{2}(D^2 b * D_\Gamma u + D_\Gamma u D^2 b) \quad (9a)$$

$$V_\Gamma u = \frac{1}{2}((D^2 b u) \otimes \nabla b + \nabla b \otimes (D^2 b u)) \quad (9b)$$

$$G_\Gamma w = \frac{1}{2}((\nabla b \otimes \nabla_\Gamma w) D^2 b + D^2 b (\nabla_\Gamma w \otimes \nabla b)) \quad (9c)$$

$$S_\Gamma w = \frac{1}{2}(D_\Gamma^2 w + *D_\Gamma^2 w) \quad (9d)$$

C_Γ and V_Γ are 1st-order and 0-order operators, respectively, that in practice operate on a tangential vector u . G_Γ is a first-order operator, and S_Γ is the symmetrization of the Hessian matrix of a scalar function w (the Hessian matrix is not symmetric in the tangential calculus [6]).

3. Weak form of the model

The elastic energy \mathcal{E}_p and kinetic energy \mathcal{E}_k of the shell under Hypothesis 2.1 were derived in [5]. The coefficient of rotational inertia is denoted $\gamma = \frac{h^2}{12}$.

$$V = \left\{ \mathbf{e} \in [H^1(\Gamma)]^2 \times H^2(\Gamma) \mid e_\Gamma = w = \frac{\partial}{\partial \nu} w = 0 \text{ on } \Upsilon_0 \right\} \quad (10)$$

We wish to define the weak form of the model in terms of two symmetric bilinear forms, $\mathfrak{m}(\mathbf{e}, \hat{\mathbf{e}})$ and $\mathfrak{a}(\mathbf{e}, \hat{\mathbf{e}})$.

Theorem 2 (Weak form of the model) *The displacement of the midsurface of the shell \mathbf{e} satisfies the following equality:*

$$\int_0^\tau [-\mathfrak{m}(\partial_t \mathbf{e}, \partial_t \hat{\mathbf{e}}) + \mathfrak{a}(\mathbf{e}, \hat{\mathbf{e}})] dt = 0 \quad (11)$$

for all test functions $\hat{\mathbf{e}} \in V$, with $\mathfrak{m}(\mathbf{e}, \hat{\mathbf{e}})$ and $\mathfrak{a}(\mathbf{e}, \hat{\mathbf{e}})$ given by

$$\begin{aligned} -\mathfrak{m}(\mathbf{e}, \hat{\mathbf{e}}) &= -\rho[2(e_\Gamma, \hat{e}_\Gamma)_\Gamma + 2\gamma((D^2 b)e_\Gamma, (D^2 b)\hat{e}_\Gamma)_\Gamma - \gamma(\nabla_\Gamma w, (D^2 b)\hat{e}_\Gamma)_\Gamma \\ &\quad - \gamma((D^2 b)e_\Gamma, \nabla_\Gamma \hat{w})_\Gamma + 2(w, \hat{w})_\Gamma + 2\gamma(\nabla_\Gamma w, \nabla_\Gamma \hat{w})_\Gamma] \end{aligned} \quad (12)$$

$$\begin{aligned} \mathfrak{a}(\mathbf{e}, \hat{\mathbf{e}}) &= 2\lambda\gamma(\Delta_\Gamma w, \Delta_\Gamma \hat{w})_\Gamma + 4\mu\gamma(w, \operatorname{tr}(S_\Gamma \hat{w}(D^2 b)^2))_\Gamma \\ &\quad + 4\mu\gamma(\operatorname{tr}(S_\Gamma w(D^2 b)^2), \hat{w})_\Gamma + 4\lambda(Hw, \operatorname{div}_\Gamma \hat{e}_\Gamma)_\Gamma \\ &\quad + 4\lambda(\operatorname{div}_\Gamma e_\Gamma, H\hat{w})_\Gamma + 4\mu(w, \operatorname{tr}(\varepsilon_\Gamma(\hat{e}_\Gamma) D^2 b))_\Gamma \\ &\quad - 4\mu\gamma(\operatorname{tr}(D^3 b e_\Gamma(D^2 b)^2), \hat{w})_\Gamma + 4\mu(\operatorname{tr}(\varepsilon_\Gamma(e_\Gamma) D^2 b), \hat{w})_\Gamma \\ &\quad + 2(\sqrt{k_\gamma}w, \sqrt{k_\gamma}\hat{w})_\Gamma + 2\lambda\gamma(\operatorname{div}_\Gamma(D^2 b e_\Gamma), \operatorname{div}_\Gamma(D^2 b \hat{e}_\Gamma))_\Gamma \\ &\quad - 2\mu(D^2 b e_\Gamma, D^2 b \hat{e}_\Gamma)_\Gamma - 2\lambda\gamma(\operatorname{div}_\Gamma(D^2 b e_\Gamma), \operatorname{tr}(C_\Gamma \hat{e}_\Gamma))_\Gamma \\ &\quad - 2\lambda\gamma(\operatorname{tr}(C_\Gamma e_\Gamma), \operatorname{div}_\Gamma(D^2 b \hat{e}_\Gamma))_\Gamma + 2\lambda\gamma(\operatorname{tr}(C_\Gamma e_\Gamma), \operatorname{tr}(C_\Gamma \hat{e}_\Gamma))_\Gamma \\ &\quad - 2\lambda\gamma(\operatorname{tr}(D^3 b e_\Gamma), \Delta_\Gamma \hat{w})_\Gamma - 2\lambda\gamma(\Delta_\Gamma w, \operatorname{tr}(D^3 b \hat{e}_\Gamma))_\Gamma \\ &\quad - 2\lambda\gamma((4H^2 - 2K)w, \operatorname{tr}(D^3 b \hat{e}_\Gamma))_\Gamma - 2\lambda\gamma(\operatorname{tr}(D^3 b e_\Gamma), (4H^2 - 2K)\hat{w})_\Gamma \\ &\quad + 2\lambda(\operatorname{div}_\Gamma e_\Gamma, \operatorname{div}_\Gamma \hat{e}_\Gamma)_\Gamma + 2\lambda\gamma((4H^2 - 2K)w, \Delta_\Gamma \hat{w})_\Gamma \\ &\quad + 2\lambda\gamma(\Delta_\Gamma w, (4H^2 - 2K)\hat{w})_\Gamma - 4\mu\gamma(w, \operatorname{tr}(D^3 b \hat{e}_\Gamma(D^2 b)^2))_\Gamma \\ &\quad - 4\mu\gamma \int_\Gamma \operatorname{tr}(\varepsilon_\Gamma(D^2 b e_\Gamma) S_\Gamma \hat{w}) + 4\mu \int_\Gamma \operatorname{tr}(\varepsilon_\Gamma(e_\Gamma) \varepsilon_\Gamma(\hat{e}_\Gamma)) \\ &\quad + 4\mu\gamma \int_\Gamma \operatorname{tr}(\varepsilon_\Gamma(D^2 b e_\Gamma) \varepsilon_\Gamma(D^2 b \hat{e}_\Gamma)) + 4\mu\gamma \int_\Gamma \operatorname{tr}(C_\Gamma e_\Gamma C_\Gamma \hat{e}_\Gamma) \\ &\quad - 4\mu\gamma \int_\Gamma \operatorname{tr}(C_\Gamma e_\Gamma \varepsilon_\Gamma(D^2 b \hat{e}_\Gamma)) - 4\mu\gamma \int_\Gamma \operatorname{tr}(\varepsilon_\Gamma(D^2 b e_\Gamma) C_\Gamma \hat{e}_\Gamma) \end{aligned}$$

$$\begin{aligned}
& + 4\mu\gamma \int_{\Gamma} \text{tr}(C_{\Gamma} e_{\Gamma} S_{\Gamma} \hat{w}) + 4\mu\gamma \int_{\Gamma} \text{tr}(S_{\Gamma} w C_{\Gamma} e_{\Gamma}) \\
& + 4\mu\gamma \int_{\Gamma} \text{tr}((S_{\Gamma} w + G_{\Gamma} w)(S_{\Gamma} \hat{w} + G_{\Gamma} \hat{w})) - 4\mu\gamma \int_{\Gamma} \text{tr}(S_{\Gamma} w \varepsilon_{\Gamma} (D^2 b e_{\Gamma})) \\
\end{aligned} \tag{13}$$

Proof. Let us consider a final time τ . Among all kinematically admissible displacements, the actual motion of the shell will make stationary the Lagrangian $\mathcal{L}(\mathbf{e}) = \int_0^\tau \mathcal{E}_k(\mathbf{e}) - \mathcal{E}_p(\mathbf{e})$. We consider the Gâteaux-derivative in a direction $\hat{\mathbf{e}}$,

$$\forall \hat{\mathbf{e}}, \left. \frac{\partial \mathcal{L}(\mathbf{e} + \theta \hat{\mathbf{e}})}{\partial \theta} \right|_{\theta=0} = 0 \tag{14}$$

with $\hat{\mathbf{e}}(0) = \partial_t \hat{\mathbf{e}}(0) = \hat{\mathbf{e}}(\tau) = \partial_t \hat{\mathbf{e}}(\tau) = 0$. Simplifying and using and using various properties of the operators defined in (9) gives $\alpha(\mathbf{e}, \hat{\mathbf{e}})$ in the required form. The constant k_{γ} (a positive real number which depends on the curvature, the Lamé coefficients, and the rotational inertia coefficient) is defined by

$$\begin{aligned}
k_{\gamma} = & 4H^2\lambda + (8H^2 - 4K)\mu + \gamma((4H^2 - 2K)^2\lambda \\
& + 2(16H^2 - 16H^2K + 2K^2)\mu) > 0
\end{aligned}$$

■

Proposition 3 (Ellipticity of the strain operator) *The bilinear form $\alpha(\mathbf{e}, \mathbf{e})$ defined in (13) is elliptic on V where the space V is defined by (10).*

Proof. It is shown in [5] that the strain tensor presented in (8) is identical to the Koiter shell model with ‘modified’ change of curvature tensor. Thus, as long as the shell is clamped on part of the boundary, the proof of Proposition 3 follows directly from the work of Bernadou and Ciarlet[3] (see also [4], and [1] pp. 23-30). ■

Proposition 4 (Existence, uniqueness of weak solutions) *Let Γ be a bounded open set with boundary Υ as previously described. Then there exists a unique solution $\mathbf{e} \in C([0, \infty); V)$ to the problem (11), with the space V given by (10).*

Proof. The proof follows immediately from Proposition 3 and the Lax-Milgram theorem. ■

4. Strong form of the model

Here we wish to integrate the symmetric bilinear forms given in Theorem 2 in order to derive a strong form of the model and the associated boundary conditions on the free part of the boundary Υ_1 . Essentially, we wish to define two operators, \mathcal{M}_γ and \mathcal{A} , so that \mathcal{M}_γ corresponds to the kinetic energy of the model and \mathcal{A} to the potential energy. Integration will give the strong form of \mathcal{M}_γ , \mathcal{A} . In [5] this calculation is done assuming a test function $\hat{\mathbf{e}}$ which is identically zero on the boundary Υ , as a shell with fully clamped boundary was considered. By contrast here we consider a shell which is clamped on only part of the boundary, Υ_0 , and is free on Υ_1 .

Theorem 5 (Strong form of the model) Define the following operator \mathcal{C} acting on a matrix A :

$$\mathcal{C}(A) = \lambda \operatorname{tr}(A) I + 2\mu A \quad (15)$$

the expression $\tilde{\chi}$

$$\tilde{\chi} = C_{\Gamma} e_{\Gamma} - \varepsilon_{\Gamma} (D^2 b e_{\Gamma}) \quad (16)$$

and the parameter $\beta = (\lambda + 2\mu)^{-1}$. Then, $\mathbf{e} \in C([0, \infty); V)$ satisfies the following system of shell equations which holds on $\Gamma \times (0, \infty)$:

$$\begin{aligned} \partial_{tt} w - \gamma \Delta_{\Gamma} \partial_{tt} w + \Delta_{\Gamma}^2 w + \frac{\gamma}{2} \operatorname{div}_{\Gamma} (D^2 b \partial_{tt} e_{\Gamma}) + P_1(e_{\Gamma}) + Q_1(w) &= 0 \\ (I + \gamma(D^2 b)^2) \partial_{tt} e_{\Gamma} - \beta [\gamma^{-1} \operatorname{div}_{\Gamma} \mathcal{C}(\varepsilon_{\Gamma}(e_{\Gamma})) + D^2 b \operatorname{div}_{\Gamma} \mathcal{C}(\tilde{\chi}) \\ - \operatorname{div}_{\Gamma} (D^2 b \mathcal{C}(\tilde{\chi}))] - \frac{\gamma}{2} D^2 b \nabla_{\Gamma} \partial_{tt} w + P_2(w) + Q_2(e_{\Gamma}) &= 0 \end{aligned}$$

where P_1 denotes coupling terms and Q_1 denotes lower order terms in the plate equation; and P_2 , Q_2 in the wave equation:

$$\begin{aligned} P_1(e_{\Gamma}) &= \beta [2\lambda H \gamma^{-1} \operatorname{div}_{\Gamma} e_{\Gamma} + 2\mu \gamma^{-1} \operatorname{tr}(D^2 b \varepsilon_{\Gamma}(e_{\Gamma})) \\ &\quad + 2\mu \operatorname{div}_{\Gamma} \operatorname{div}_{\Gamma}(\tilde{\chi}) + 4\mu H \operatorname{tr}(D^3 b e_{\Gamma} D^2 b) \\ &\quad - \lambda \Delta_{\Gamma} \langle 2\nabla_{\Gamma} H, e_{\Gamma} \rangle - \lambda(4H^2 - 2K) \langle 2\nabla_{\Gamma} H, e_{\Gamma} \rangle] \\ &\quad - 2\mu \langle (D^2 b)^2 .. D^3 b, e_{\Gamma} \rangle \end{aligned} \quad (17a)$$

$$\begin{aligned} Q_1(w) &= \beta [k_{\gamma} w + 4\mu \operatorname{div}_{\Gamma} ((D^2 b)^2 \nabla_{\Gamma} w) + \lambda \Delta_{\Gamma} ((4H^2 - 2K) w) \\ &\quad + 2\mu \operatorname{div}_{\Gamma} \operatorname{div}_{\Gamma} ((D^2 b)^2 w) + \lambda(4H^2 - 2K) \Delta_{\Gamma} w \\ &\quad + 2\mu \operatorname{div}_{\Gamma} (K \nabla_{\Gamma} w)] + 2\mu \operatorname{tr}(S_{\Gamma} w (D^2 b)^2) \\ &\quad + 4\mu H \operatorname{tr}((D^2 b)^3 w) \end{aligned} \quad (17b)$$

$$\begin{aligned}
P_2(w) = & \beta[-2\lambda\gamma^{-1}\nabla_\Gamma(Hw) + 2\mu\gamma^{-1}\operatorname{div}_\Gamma(w D^2b) \\
& + \lambda 2\nabla_\Gamma H(\Delta_\Gamma w - (4H^2 - 2K)w) - 2\mu(D^2b)^2 .. D^3bw] \\
& - 2\mu\operatorname{div}_\Gamma(D^2b S_\Gamma w) + 2\mu D^2b \operatorname{div}_\Gamma(S_\Gamma w)
\end{aligned} \tag{17c}$$

$$Q_2(e_\Gamma) = -2\beta\mu\gamma^{-1}(K e_\Gamma + 2(D^2b)^2 e_\Gamma) \tag{17d}$$

the following free boundary conditions on $\Upsilon_1 \times (0, \infty)$:

$$\mathcal{C}(w D^2b + \varepsilon_\Gamma(e_\Gamma)) \cdot \nu = 0$$

$$\begin{aligned}
& \langle (\lambda\beta\nabla_\Gamma \operatorname{tr}(D^3b e_\Gamma) + 4\mu\beta(D^2b)^2 \nabla_\Gamma w + 2\mu\beta K \nabla_\Gamma w - \nabla_\Gamma(\Delta_\Gamma w), \nu) \\
& + \langle \gamma(D^2b \partial_{tt}e_\Gamma - 2\nabla_\Gamma \partial_{tt}w) + 2\mu\beta B_\Gamma^3(C_\Gamma e_\Gamma - \varepsilon_\Gamma(D^2b e_\Gamma) \\
& + (D^2b)^2 w), \nu \rangle - \beta\lambda \langle \nabla_\Gamma((4H^2 - 2K)w), \nu \rangle + 2\mu B_\Gamma^2 w = 0
\end{aligned}$$

$$\begin{aligned}
& \lambda\beta \operatorname{tr}(D^3b e_\Gamma) + \lambda\beta(4H^2 - 2K)w - \Delta_\Gamma w \\
& + 2\mu\beta \langle C_\Gamma e_\Gamma - \varepsilon_\Gamma(D^2b e_\Gamma) + (D^2b)^2 w, \nu \rangle + 2\mu\beta B_\Gamma^1 w = 0
\end{aligned}$$

and clamped boundary conditions on $\Upsilon_0 \times (0, \infty)$:

$$w = \frac{\partial}{\partial \nu} w = e_\Gamma = 0 \tag{18}$$

Proof. We search for the explicit representations of the following Green's-type formulas:

$$\begin{aligned}
& \int_0^\tau -\mathfrak{m}(\partial_t \mathbf{e}, \partial_t \hat{\mathbf{e}}) dt = \\
& \int_0^\tau \int_\Gamma \langle \mathcal{M}_\gamma \partial_{tt} \mathbf{e}, \hat{\mathbf{e}} \rangle dtd\Gamma - \int_0^\tau \int_{\Upsilon_1} \mathfrak{B}_{\mathcal{M}}(\partial_{tt} \mathbf{e}, \hat{\mathbf{e}}) dtd\Upsilon_1
\end{aligned} \tag{19}$$

$$(\mathcal{A}\mathbf{e}, \hat{\mathbf{e}})_\Gamma = \int_\Gamma \langle \mathcal{A}\mathbf{e}, \hat{\mathbf{e}} \rangle d\Gamma = \int_{\Upsilon_1} \mathfrak{B}_{\mathcal{A}}(\mathbf{e}, \hat{\mathbf{e}}) d\Upsilon_1 + \mathfrak{a}(\mathbf{e}, \hat{\mathbf{e}}) \tag{20}$$

where $\mathfrak{B}(\mathbf{e}, \hat{\mathbf{e}})$ are expressions containing boundary terms only. We first note that the regularity of the weak solutions is high enough to permit the necessary integration by parts to derive the strong form.

Proposition 6 Let Γ with boundary Υ be as previously described and assume $\mathbf{e} \in C((0, \infty); V)$ such that

$$\int_0^\tau [-\mathfrak{m}(\partial_t \mathbf{e}, \partial_t \hat{\mathbf{e}}) + \mathfrak{a}(\mathbf{e}, \hat{\mathbf{e}})] dt = 0 \quad \forall \hat{\mathbf{e}} \in V \tag{21}$$

with the space V defined in (10). Then $\mathbf{e} \in C((0, \infty); V_1)$ where

$$V_1 = \left\{ \mathbf{e} \in [H^2(\Gamma)]^2 \times H^4(\Gamma) \mid e_\Gamma = w = \frac{\partial}{\partial \nu} w = 0 \text{ on } \Upsilon_0 \right\} \quad (22)$$

and the restriction of \mathbf{e} to the boundary $\mathbf{e}|_\Upsilon \in [H^{1/2}(\Upsilon)]^2 \times H^{3/2}(\Upsilon)$.

Proof. Choose $\hat{\mathbf{e}} \in [\mathcal{D}(\Gamma)]^3$. Then the integrals over Υ in (19) and (20) vanish so that we have

$$\int_0^\tau \int_\Gamma \langle \mathcal{M}_\gamma \partial_{tt} \mathbf{e} + \mathcal{A} \mathbf{e}, \hat{\mathbf{e}} \rangle d\Gamma dt = 0, \quad \forall \hat{\mathbf{e}} \in [\mathcal{D}(\Gamma)]^3 \quad (23)$$

Thus $\mathcal{M}_\gamma \partial_{tt} \mathbf{e} + \mathcal{A} \mathbf{e} = 0$ in the sense of distributions. This equality also holds in $[L_2(\Gamma)]^3$, since $[\mathcal{D}(\Gamma)]^3$ is dense in $[L_2(\Gamma)]^3$ and the Γ integrals are continuous with respect to $\hat{\mathbf{e}}$ in the $L_2(\Gamma)$ topology. Next, we consider an arbitrary $\hat{\mathbf{e}}$ and integrate (21), giving

$$\int_0^\tau \int_\Gamma \langle \mathcal{M}_\gamma \partial_{tt} \mathbf{e} + \mathcal{A} \mathbf{e}, \hat{\mathbf{e}} \rangle d\Gamma dt - \int_0^\tau \int_\Upsilon \mathfrak{B}_\mathcal{M}(\partial_{tt} \mathbf{e}, \hat{\mathbf{e}}) + \mathfrak{B}_\mathcal{A}(\mathbf{e}, \hat{\mathbf{e}}) d\Upsilon dt = 0 \quad (24)$$

By the previous step this gives immediately that

$$\int_0^\tau \int_\Upsilon \mathfrak{B}_\mathcal{M}(\partial_{tt} \mathbf{e}, \hat{\mathbf{e}}) + \mathfrak{B}_\mathcal{A}(\mathbf{e}, \hat{\mathbf{e}}) d\Upsilon_1 dt = 0 \quad (25)$$

for any $\hat{\mathbf{e}} \in [H^1(\Gamma)]^2 \times H^2(\Gamma)$. However, for any $\hat{\mathbf{e}}|_\Upsilon \in [H^{1/2}(\Upsilon)]^2 \times H^{3/2}(\Upsilon)$, by surjectivity of the trace map there exists an extension $\tilde{\mathbf{e}} \in [H^1(\Gamma)]^2 \times H^2(\Gamma)$ such that $\tilde{\mathbf{e}}|_\Upsilon = \hat{\mathbf{e}}|_\Upsilon$. This gives that (25) holds true for all $\hat{\mathbf{e}}|_\Upsilon \in [H^{1/2}(\Upsilon)]^2 \times H^{3/2}(\Upsilon)$, and thus again by density we have that $\mathfrak{B}_\mathcal{M} \partial_{tt} \mathbf{e} + \mathfrak{B}_\mathcal{A} \mathbf{e} = 0$ in $[L^2(\Upsilon)]^3$. ■

With respect to the kinetic energy, integration by parts in time followed by an application of the Green's formula (4) gives that

$$\begin{aligned} (\mathcal{M}_\gamma \partial_{tt} \mathbf{e}, \hat{\mathbf{e}})_\Gamma &= \rho [(2\partial_{tt} e_\Gamma + 2\gamma(D^2 b)^2 \partial_{tt} e_\Gamma, \hat{e}_\Gamma)_\Gamma \\ &\quad - \gamma((D^2 b) \nabla_\Gamma \partial_{tt} w, \hat{e}_\Gamma)_\Gamma + \gamma(\operatorname{div}_\Gamma((D^2 b) \partial_{tt} e_\Gamma), \hat{w})_\Gamma \\ &\quad + (2\partial_{tt} w - 2\gamma \Delta_\Gamma \partial_{tt} w, \hat{w})_\Gamma] \end{aligned} \quad (26)$$

$$\int_{\Upsilon_1} \mathfrak{B}_\mathcal{M}(\partial_{tt} \mathbf{e}, \hat{\mathbf{e}}) d\Upsilon_1 = -\rho \gamma \int_{\Upsilon_1} ((D^2 b) \partial_{tt} e_\Gamma \cdot \nu - 2\nabla_\Gamma \partial_{tt} w \cdot \nu) \hat{w} d\Upsilon_1 \quad (27)$$

We note that terms involving the principal curvature H coming from (4) are zero since both $\nabla_\Gamma \hat{w}$ and $D^2 b \hat{e}_\Gamma$ are tangential vectors. Now we

need to integrate each of the thirty-two terms of (13) to derive a strong representation of the operator \mathcal{A} . Space considerations do not allow us to show all of these calculations here, thus we note only the procedure followed, especially that with the highest-order terms. The first term of (13) will yield the required tangential biharmonic operator Δ_Γ^2 in the strong form. However, the second term of this equation is *also* a fourth-order operator, and analogously with the plate case we desire to combine the two terms $\Delta_\Gamma w \Delta_\Gamma \hat{w}$ and $\text{tr}(S_\Gamma w S_\Gamma \hat{w})$. The problem here lies in the fact that the matrix $D_\Gamma^2 w$ is *not* symmetric and is not equivalent to the restriction of the Hessian matrix of the canonical extension $w \circ p$. In fact the two terms differ by a first-derivative correction, as is seen in the following identity [6]: $D_\Gamma^2 w - (D^2 b \nabla_\Gamma w) \otimes \nabla b = D^2(w \circ p)|_\Gamma$. We use this identity and the definitions of the operators S_Γ and G_Γ to write that

$$2\lambda\gamma(\Delta_\Gamma w, \Delta_\Gamma \hat{w})_\Gamma + 4\mu\gamma \int_\Gamma \text{tr}((S_\Gamma w + G_\Gamma w)(S_\Gamma \hat{w} + G_\Gamma \hat{w})) d\Gamma = \\ 2\gamma(\lambda + 2\mu)(\Delta_\Gamma w, \Delta_\Gamma \hat{w})_\Gamma - 8\mu\gamma(D^2 b \nabla_\Gamma w, D^2 b \nabla_\Gamma \hat{w})_\Gamma + 4\mu\gamma \int_\Gamma \text{a.t. } d\Gamma$$

where *a.t.* stands for additional terms. Analogous to the plate case we are able to convert the additional term in this equation to an integral over the boundary Υ by showing that it is the divergence of an appropriate vector. In fact the last integral is equivalent to the following:

$$4\mu\gamma \int_\Upsilon [B_\Gamma^1(D^2(w \circ p)|_\Gamma) \nabla_\Gamma \hat{w} \cdot \nu - \hat{w} B_\Gamma^2(D^2(w \circ p)|_\Gamma)] d\Upsilon \quad (28)$$

where

$$B_\Gamma^1 A = -(\tau \otimes \tau) .. A \quad (29a)$$

$$B_\Gamma^2 A = \langle \nabla_\Gamma ((\tau \otimes \nu) .. A), \tau \rangle \quad (29b)$$

with ν and τ being the normal and tangent vectors to the boundary Υ . Denoting $\nu = (\nu_1, \nu_2)$ and $\tau = (-\nu_2, \nu_1)$ and expanding equations (29) shows that the $B_\Gamma^1(D^2(w \circ p)|_\Gamma)$ and $B_\Gamma^2(D^2(w \circ p)|_\Gamma)$ in (28) are exactly the restriction to Γ of the standard B_1, B_2 operators (which appear in the modeling of Kirchhoff plates) operating on the canonical extension $w \circ p$ (see, e.g., [14], [15]). Integration of this term and the others follow through explicit use of (4), (6) and (5). For readability we define also

$$B_\Gamma^3 A = B_\Gamma^2 A + \langle \text{div}_\Gamma A, \nu \rangle \quad (29c)$$

and the following operator \mathcal{C} acting on a matrix A : $\mathcal{C}(A) = \lambda \text{tr}(A) I + 2\mu A$. Finally, we make a change of time variable $t \rightarrow t \sqrt{\frac{\gamma(\lambda+2\mu)}{\rho}}$ in both equations and divide both equations by the term $2\gamma(\lambda + 2\mu)$. Defining $\beta = (\lambda + 2\mu)^{-1}$ gives the final form of the equations. ■

Theorem 7 (Existence and uniqueness of strong solutions) *Let Γ be a bounded open set with boundary Υ as previously described. Then there exists a unique solution $\mathbf{e} \in C((0, \infty); V)$ to the system of equations presented in Theorem 4 , with the space V given by (10).*

Proof. The proof follows immediately from the Lumer-Phillips theorem once Proposition 3 is established. ■

References

- [1] M. Bernadou. *Finite Element Methods for Thin Shell Problems*. Wiley and Sons, 1996.
- [2] M. Bernadou and J. M. Boisserie. *The Finite Element Method for Thin Shell Problems*. Birkhauser, Boston, 1982.
- [3] M. Bernadou and P. G. Ciarlet. Sur l'ellipticité du modèle linéaire de coques de W. T. Koiter. In *Computing Methods in Applied Sciences and Engineering, Lecture Notes in Economics and Mathematical Systems*, vol. 34, pp. 89–136. Springer-Verlag, 1976.
- [4] M. Bernadou, P. G. Ciarlet, and B. Miara. Existence theorems for two-dimensional linear shell theories. *J. Elasticity*, 34:111–138, 1994.
- [5] J. Cagnol, I. Lasiecka, C. Lebiedzik, and J.-P. Zolésio. Uniform stability in structural acoustic model with flexible curved walls. *J.Dif.Eqns.*, 186(1), 88–121, 2002.
- [6] M. Delfour and J. P. Zolésio. *Intrinsic differential geometry and theory of thin shells*. to appear, 2002.
- [7] M. C. Delfour and J.-P. Zolésio. A boundary differential equation for thin shells. *Journal of Differential Equations*, 119(2):426–449, 1995.
- [8] M. C. Delfour and J.-P. Zolésio. Differential equations for linear shells: comparison between intrinsic and classical models. In *Advances in mathematical sciences: CRM's 25 years CRM Proc. Lecture Notes*, vol. 11, pp. 481–491. AMS 1997.
- [9] P. Germain. *Mecanique*, vol. 1. Ecole Polytechnique, 1986.
- [10] G. Kirchhoff. *Vorlesungen über Mathematische Physik. Mechanik*. Leipzig, 1876.
- [11] W. T. Koiter. A consistent first approximation in the general theory of thin shells. In *Proc. Symp. on Theory of Thin Elastic Shells*, pp. 12–33. held in Delft, August 1959.
- [12] W. T. Koiter. On the foundation of the linear theory of thin elastic shells. *Proc. Kon. Nederl. Akad. Wetensch.*, B73:169–195, 1970.
- [13] J. Lagnese. *Boundary Stabilization of Thin Plates*. SIAM, Philadelphia, PA, 1989.
- [14] J. Lagnese and J.-L. Lions. *Modelling analysis and control of thin plates*, vol. 6 of *Research in Applied Mathematics*. Masson, Paris, 1988.
- [15] I. Lasiecka and R. Triggiani. *Control Theory for Partial Differential Equations*, vol. I and II. *Encyclopedia of Mathematics and its Applications*. Cambridge U. Press, 1999.
- [16] A. E. H. Love. *The Mathematical Theory of Elasticity*. Cambridge Univ. Press, 1934.

ERROR ESTIMATES FOR LINEAR-QUADRATIC ELLIPTIC CONTROL PROBLEMS

Eduardo Casas

*Departamento de Matemática Aplicada y Ciencias de la Computación
Universidad de Cantabria
39005 Santander, Spain.
eduardo.casas@unican.es*

Fredi Tröltzsch

*Technische Universität Berlin
Institut für Mathematik, Sekretariat MA 4-5
Str. d. 17. Juni 136
D-10623 Berlin, Germany
troeltzsch@math.tu-berlin.de*

Abstract The discretization of control functions by piecewise constant and piecewise linear functions is considered for linear-quadratic elliptic optimal control problems. Error estimates are derived for the optimal controls. Special emphasis is laid on the case of boundary control and convex polygonal domains.

Keywords: Optimal control, elliptic equation, error estimate

Introduction

In this paper, we discuss the error analysis for numerical approximations of the problem (P) to minimize the objective functional

$$\lambda_\Omega \|y - y_\Omega\|_{L^2(\Omega)}^2 + \lambda_\Gamma \|y - y_\Gamma\|_{L^2(\Gamma)}^2 + \lambda_1 \|u_1\|_{L^2(\Omega)}^2 + \lambda_2 \|u_2\|_{L^2(\Gamma)}^2$$

subject to the elliptic boundary value problem

$$\begin{aligned} A y &= b_1 u_1 && \text{in } \Omega \\ \partial_\nu y + \beta y &= b_2 u_2 && \text{on } \Gamma \end{aligned}$$

and to pointwise control constraints $u_a^1 \leq u_1(x) \leq u_b^1$, $u_a^2 \leq u_2(x) \leq u_b^2$. Here, a domain $\Omega \subset \mathbb{R}^N$ with boundary Γ , $N \geq 2$, real constants $u_a^i \leq u_b^i$, $i = 1, 2$, functions $b_1 \in C^{0,1}(\bar{\Omega})$, β and $b_2 \in C^{0,1}(\Gamma)$, $y_\Omega \in L^2(\Omega)$, $y_\Gamma \in L^2(\Gamma)$, and certain nonnegative constants in the objective functional are given that partially can be zero. Concerning the smoothness of Γ , if not stated otherwise, we shall work with

(A1) Ω is bounded with boundary Γ of class $C^{0,1}$.

Moreover, an elliptic differential operator A in divergence form,

$$Ay(x) = -\sum_{i,j=1}^N D_i(a_{ij}(x) D_j y(x)) + a_0(x)y(x) \quad \text{with coefficients } a_{ij}, a_0 \in L^\infty(\Omega)$$

is given. Its formally adjoint operator is denoted by A^* . The a_{ij} are assumed to satisfy the condition of uniform ellipticity

$$\sum_{i,j=1}^N \xi_i \xi_j a_{ij}(x) \geq \alpha_0 |\xi|^2 \quad \text{for all } x \in \Omega \text{ and all } \xi \in \mathbb{R}^N, \text{ where } \alpha_0 \text{ is a positive constant.}$$

By ∂_{ν_A} we denote the co-normal derivative at Γ w.r. to A .

Error estimates for elliptic control problems have already been studied by several authors for linear and nonlinear equations and distributed control. We mention [6], [7], [2], [1], and [4]. Here, we consider the case of boundary control, which is more difficult in several aspects. However, to see better the difference to distributed control, we also discuss this case.

We are able to handle also a more general problem containing terms $\|\omega y - y_\Omega\|^2$, $\|\gamma y - y_\Gamma\|^2$ with Lipschitz functions ω , γ in the objective and $b_1 u_1 + f_1$, $b_2 u_2 + f_2$ in the right hand sides of the elliptic boundary value problem. This class covers linear-quadratic sub-problems in Lagrange-Newton-SQP methods for elliptic equations, where an error analysis is desirable. The discussion of this more general problem is analogous to that for (P) but notationally more complex. Therefore, we consider the simpler problem (P).

The error analysis is performed in a, perhaps, nonstandard way. Here, we concentrate on the problem of discretizing only the control functions while leaving the elliptic equation unchanged. Equipped with these estimates, in a second step the approximation of the elliptic equation by numerical schemes such as finite element methods can be studied – then for controls restricted to an admissible set of discretized functions. The presentation of both types of estimates would exceed the size of the paper. We only briefly comment the application of FEM in the last section. The different types of controls and observations will be discussed separately. It is easy to deduce error estimates for the general problem (P)

from the particular cases. To unify the presentation, all controls will be denoted by u , and b stands for the b_i . Moreover, we delete the index i in the bounds u_a^i and u_b^i .

1. Approximation of controls by step functions

1.1. Distributed control

Distributed observation. We consider first the following particular case of (P) with $b := b_1$ and $\lambda > 0$,

$$(P^1) \quad \min J(y, u) = \|y - y_\Omega\|_{L^2(\Omega)}^2 + \lambda \|u\|_{L^2(\Omega)}^2$$

subject to

$$\begin{aligned} A y &= b u \\ \partial_{\nu_A} y + \beta y &= 0 \end{aligned} \tag{1.1}$$

and to

$$u_a \leq u(x) \leq u_b.$$

The state y is defined in $H^1(\Omega)$ as weak solution of (1.1), and the control u is considered as a function of $L^2(\Omega)$, although the constraints yield even $u \in L^\infty(\Omega)$. We assume

(A2) *The functions a_0 and β are nonnegative. At least one of the functions is not identically zero in the sense of L^∞ .*

It is known that (A1) and (A2) guarantee existence and uniqueness of $y = y(u)$ of (1.1) for arbitrary $u \in L^2(\Omega)$. Moreover, the control-to-state mapping $S : u \mapsto y$ is continuous from $L^2(\Omega)$ to $H^1(\Omega)$. We refer, for instance, to [3].

We shall consider S as mapping from $L^2(\Omega)$ to $L^2(\Omega)$, where it is continuous as well. In view of this definition, (P^1) admits the form

$$(P^1) \quad \min \|S u - y_\Omega\|_{L^2(\Omega)}^2 + \lambda \|u\|_{L^2(\Omega)}^2, \quad u \in U^{ad},$$

where $U^{ad} = \{u \in L^2(\Omega) \mid u_a \leq u(x) \leq u_b \text{ a.e. in } \Omega\}$.

Theorem 1.1 *If (A1) and (A2) are satisfied, then Problem (P^1) has a unique solution $\bar{u} \in U^{ad}$ with associated optimal state $\bar{y} = y(\bar{u})$.*

The proof of this theorem is standard. Next we state the necessary (and by convexity also sufficient) optimality conditions for \bar{u} , which are standard as well. The *variational inequality*

$$(S^*(S\bar{u} - y_\Omega) + \lambda\bar{u}, u - \bar{u})_{L^2(\Omega)} \geq 0 \quad \forall u \in U^{ad} \tag{1.2}$$

must be fulfilled. We find $S^*(S\bar{u} - y_\Omega) = S^*(\bar{y} - y_\Omega) = b\bar{p}$, where the function \bar{p} is the *adjoint state* and solves the *adjoint equation*

$$A^*\bar{p} = \bar{y} - y_\Omega, \quad \partial_{\nu_{A^*}}\bar{p} + \beta\bar{p} = 0. \tag{1.3}$$

From (1.2), the well-known projection formula

$$\bar{u}(x) = \text{Proj}_{[u_a, u_b]} \left\{ -\frac{1}{\lambda} b(x) \bar{p}(x) \right\} \quad (1.4)$$

is obtained. In this formula, $\text{Proj}_{[u_a, u_b]}$ denotes the projection mapping from \mathbb{R} onto $[u_a, u_b]$.

Lemma 1.1 *The optimal control \bar{u} of (P^1) has the regularity $\bar{u} \in H^1(\Omega)$.*

Proof: From $\bar{y} - y_\Omega \in L^2(\Omega)$ we obtain $\bar{p} \in H^1(\Omega)$ for the adjoint state. We also have $b \bar{p} \in H^1(\Omega)$, since $b \in C^{0,1}(\bar{\Omega})$, see [8], Thm. 1.4.1.1. The projection operator $y(\cdot) \mapsto \text{Proj}_{[u_a, u_b]} y(\cdot)$ is continuous in $H^1(\Omega)$. This follows from the continuity of the operator $y(\cdot) \mapsto |y(\cdot)|$ in $H^1(\Omega)$, see [10] or [3], Appendix. Therefore, (1.4) yields $\bar{u} \in H^1(\Omega)$. ■

Next we introduce the approximation of the control function u by step functions. We assume that $\bar{\Omega} = \cup_{j=1}^m \bar{\Omega}_j$, where $\Omega_j \subset \Omega$ are finitely many pairwise disjoint (open) subdomains such that $\text{diam}(\Omega_j) \leq \sigma \forall j \in \{1, \dots, m\}$. The variable σ can be considered as the mesh-size of an associated grid, for instance, a partition by triangles or rectangles, if $\Omega \in \mathbb{R}^2$.

Let $\Pi_\sigma : H^1(\Omega) \rightarrow L^2(\Omega)$ denote the L^2 -projection operator onto the space of step functions defined by

$$\Pi_\sigma u(x) = \frac{1}{|\Omega_j|} \int_{\Omega_j} u(\xi) d\xi, \quad x \in \Omega_j,$$

$j = 1, \dots, m$, where $|\Omega_j|$ denotes the Lebesgue measure of Ω_j . There exists a constant c_π such that

$$\|\Pi_\sigma u - u\|_{L^2(\Omega)} \leq c_\pi \sigma \|u\|_{H^1(\Omega)} \quad (1.5)$$

holds for all $u \in H^1(\Omega)$, [5], chpt. II, Thm. 15.3. We introduce the admissible set of step functions

$$U_\sigma^{ad} = \{u \in U^{ad} \mid u(x) \text{ is constant on each } \Omega_j, j = 1, \dots, m\}.$$

The finite-dimensional approximation of (P^1) is defined by substituting U_σ^{ad} for U^{ad} ,

$$(P_\sigma^1) \quad \min \quad \|S u - y_\Omega\|_{L^2(\Omega)}^2 + \lambda \|u\|_{L^2(\Omega)}^2, \quad u \in U_\sigma^{ad}.$$

Considering U_σ^{ad} as a subset of $L^2(\Omega)$, this problem can be discussed in the same way as (P^1) . We have exactly one optimal control \bar{u}_σ in U_σ^{ad} . The associated variational inequality is

$$(S^*(S\bar{u}_\sigma - y_\Omega) + \lambda \bar{u}_\sigma, u - \bar{u}_\sigma)_{L^2(\Omega)} \geq 0 \quad \forall u \in U_\sigma^{ad}. \quad (1.6)$$

We put $\bar{y}_\sigma = y(\bar{u}_\sigma)$ and define $\bar{p}_\sigma = p(\bar{y}_\sigma)$ by

$$A^* \bar{p}_\sigma = y(\bar{u}_\sigma) - y_\Omega, \quad \partial_{\nu_{A^*}} \bar{p}_\sigma + \beta \bar{p}_\sigma = 0. \quad (1.7)$$

Notice that \bar{p}_σ is not a step function!

Theorem 1.2 *If (A1) and (A2) are satisfied, then there is a constant c_1 depending on $\|\bar{u}\|_{H^1(\Omega)}$ but not on σ and \bar{u}_σ such that*

$$\|\bar{u} - \bar{u}_\sigma\|_{L^2(\Omega)} \leq c_1 \sigma.$$

Proof: From (1.2) and (1.6) we find by inserting $u = \bar{u}_\sigma$ and $u = \Pi_\sigma \bar{u}$, respectively,

$$\begin{aligned} (S^*(S\bar{u} - y_\Omega) + \lambda \bar{u}, \bar{u}_\sigma - \bar{u})_{L^2(\Omega)} &\geq 0 \\ (S^*(S\bar{u}_\sigma - y_\Omega) + \lambda \bar{u}_\sigma, \Pi_\sigma \bar{u} - \bar{u}_\sigma)_{L^2(\Omega)} &\geq 0. \end{aligned} \quad (1.8)$$

Next, we rewrite the second inequality in (1.8) as

$$\begin{aligned} (S^*(S\bar{u}_\sigma - y_\Omega) + \lambda \bar{u}_\sigma, \bar{u} - \bar{u}_\sigma)_{L^2(\Omega)} \\ + (S^*(S\bar{u}_\sigma - y_\Omega) + \lambda \bar{u}_\sigma, \Pi_\sigma \bar{u} - \bar{u})_{L^2(\Omega)} \geq 0 \end{aligned}$$

and add the first one. Then we obtain by $S^*(S\bar{u}_\sigma - y_\Omega) = b p(\bar{y}_\sigma)$

$$\begin{aligned} (S^* S(\bar{u} - \bar{u}_\sigma), \bar{u}_\sigma - \bar{u})_{L^2(\Omega)} - \lambda \|\bar{u} - \bar{u}_\sigma\|_{L^2(\Omega)}^2 \\ + (b p(\bar{y}_\sigma) + \lambda \bar{u}_\sigma, \Pi_\sigma \bar{u} - \bar{u})_{L^2(\Omega)} \geq 0. \end{aligned}$$

In view of $(S^* S(\bar{u} - \bar{u}_\sigma), \bar{u}_\sigma - \bar{u})_{L^2(\Omega)} = -\|S(\bar{u} - \bar{u}_\sigma)\|_{L^2(\Omega)}^2$ and the known relation of orthogonality

$$(u_\sigma, \Pi_\sigma u - u)_{L^2(\Omega)} = 0 \quad \forall u \in L^2(\Omega) \quad \forall u_\sigma \in \Pi_\sigma L^2(\Omega) \quad (1.9)$$

we arrive at

$$\begin{aligned} \lambda \|\bar{u} - \bar{u}_\sigma\|_{L^2(\Omega)}^2 &\leq -\|S(\bar{u} - \bar{u}_\sigma)\|_{L^2(\Omega)}^2 + (b p(\bar{y}_\sigma), \Pi_\sigma \bar{u} - \bar{u})_{L^2(\Omega)} \\ &\leq (b p(\bar{y}_\sigma), \Pi_\sigma \bar{u} - \bar{u})_{L^2(\Omega)}. \end{aligned}$$

Now we might estimate the right hand side by the Cauchy-Schwarz inequality and take the square root. We would obtain an error estimate of the order $\sqrt{\sigma}$, which is not optimal. Instead, we continue by

$$\begin{aligned} \lambda \|\bar{u} - \bar{u}_\sigma\|_{L^2(\Omega)}^2 &\leq (b p(\bar{y}_\sigma), \Pi_\sigma \bar{u} - \bar{u})_{L^2(\Omega)} \\ &= (b p(\bar{y}_\sigma) - \Pi_\sigma(b p(\bar{y}_\sigma)), \Pi_\sigma \bar{u} - \bar{u})_{L^2(\Omega)} \\ &\quad + (\Pi_\sigma(b p(\bar{y}_\sigma)), \Pi_\sigma \bar{u} - \bar{u})_{L^2(\Omega)} \\ &\leq \|b p(\bar{y}_\sigma) - \Pi_\sigma(b p(\bar{y}_\sigma))\|_{L^2(\Omega)} \|\Pi_\sigma \bar{u} - \bar{u}\|_{L^2(\Omega)} \\ &\leq (c_\Pi \sigma)^2 \|b p(\bar{y}_\sigma)\|_{H^1(\Omega)} \|\bar{u}\|_{H^1(\Omega)}, \end{aligned}$$

where we have used (1.9) and (1.5). The norm $\|b p(\bar{y}_\sigma)\|_{H^1(\Omega)}$ still depends on σ . However, it holds with a generic constant c

$$\begin{aligned}\|b p(\bar{y}_\sigma)\|_{H^1(\Omega)} &\leq c \|p(\bar{y}_\sigma)\|_{H^1(\Omega)} \leq c (\|\bar{y}_\sigma\|_{L^2(\Omega)} + \|y_\Omega\|_{L^2(\Omega)}) \\ &\leq c (\|\bar{u}_\sigma\|_{L^2(\Omega)} + \|y_\Omega\|_{L^2(\Omega)}) \leq c\end{aligned}$$

since U_σ^{ad} is uniformly bounded. Altogether, we have obtained the result of the theorem, where c_1 depends on $\|\bar{u}\|_{H^1(\Omega)}$ ■

Boundary observation. Completely analogous we can discuss (P^1) with the objective functional

$$J(y, u) = \|y - y_\Gamma\|_{L^2(\Gamma)}^2 + \lambda \|u\|_{L^2(\Omega)}^2.$$

Here, the control-to-observation mapping S is defined by $S : L^2(\Omega) \rightarrow L^2(\Gamma)$ and $S u = y|_\Gamma$, where y is the solution to (1.1). From $y \in H^1(\Omega)$ we conclude $y|_\Gamma \in H^{1/2}(\Gamma)$. Therefore, S is well defined. The optimal quantities are denoted as before. The projection formula (1.4) for the optimal control holds as well, but the adjoint state \bar{p} is defined by

$$A^* \bar{p} = 0, \quad \partial_{\nu_A} \bar{p} + \beta \bar{p} = \bar{y}|_\Gamma - y_\Gamma.$$

The boundary data $\bar{y}|_\Gamma - y_\Gamma$ belong at least to $L^2(\Gamma)$, thus we have $\bar{p} \in H^1(\Omega)$, and (1.4) yields $\bar{u} \in H^1(\Omega)$. This is the only information we needed to prove Theorem 1.2. Exactly the same arguments show that the theorem remains true for the case of boundary observation.

1.2. Boundary control

Distributed observation. Next we discuss the problem

$$(P^2) \quad \min J(y, u) = \|y - y_\Omega\|_{L^2(\Omega)}^2 + \lambda \|u\|_{L^2(\Gamma)}^2$$

subject to $u \in U^{ad}$ and

$$\begin{aligned}A y &= 0 \\ \partial_{\nu_A} y + \beta y &= b u,\end{aligned} \tag{1.10}$$

where $U^{ad} = \{u \in L^2(\Gamma) \mid u_a \leq u(x) \leq u_b \text{ a.e. on } \Gamma\}$, $b := b_2$ and $\lambda > 0$. In this case, we must cope with a certain lack of regularity of the adjoint state. We have to construct an associated partitioning of Γ to define piecewise constant controls. To do this, we assume that $\Gamma = \cup_{j=1}^m \bar{\Gamma}_j$, where $\bar{\Gamma}_j \subset \Gamma$ are finitely many pairwise disjoint open and connected subsets of Γ such that $diam(\bar{\Gamma}_j) \leq \sigma \forall j \in \{1, \dots, m\}$. If Γ is the boundary of a two-dimensional domain Ω , this may be accomplished by an

equidistant splitting of Γ into m pieces of arclength σ . The projector Π_σ is now

$$\Pi_\sigma u(x) = \frac{1}{|\Gamma_j|} \int_{\Gamma_j} u(\xi) d\Gamma(\xi), \quad x \in \Gamma_j,$$

$j=1, \dots, m$. As in (1.5), the estimate $\|\Pi_\sigma u - u\|_{L^2(\Gamma)} \leq c_\pi \sigma \|u\|_{H^1(\Gamma)}$ holds. Here, the set of admissible piecewise constant boundary controls is

$$U_\sigma^{ad} = \{u \in U^{ad} \mid u(x) \text{ is constant on each } \Gamma_j, j = 1, \dots, m.\}$$

We perform the error analysis for the approximate problem (P_σ^2) obtained from (P^2) by substituting U_σ^{ad} for U^{ad} . The control-to-state mapping S is now defined by $Su = y$, $S : L^2(\Gamma) \rightarrow L^2(\Omega)$, where y is the solution to (1.10). The adjoint operator S^* maps $L^2(\Omega)$ into $L^2(\Gamma)$. We denote again by \bar{u} , \bar{u}_σ , $\bar{y} = y(\bar{u})$, and $\bar{y}_\sigma = y(\bar{u}_\sigma)$ the optimal solutions and define the adjoint state \bar{p} by the adjoint equation

$$A^* \bar{p} = \bar{y} - y_\Omega, \quad \partial_{\nu_{A^*}} \bar{p} + \beta \bar{p} = 0. \quad (1.11)$$

The optimality conditions yield the two inequalities

$$\begin{aligned} (S^*(S\bar{u} - y_\Omega) + \lambda \bar{u}, \bar{u}_\sigma - \bar{u})_{L^2(\Gamma)} &\geq 0 \\ (S^*(S\bar{u}_\sigma - y_\Omega) + \lambda \bar{u}_\sigma, \Pi_\sigma \bar{u} - \bar{u}_\sigma)_{L^2(\Gamma)} &\geq 0 \end{aligned}$$

together with the projection formula

$$\bar{u}(x) = Proj_{[u_a, u_b]} \left\{ -\frac{1}{\lambda} b(x) \bar{p}(x)|_\Gamma \right\}. \quad (1.12)$$

Now, the standard H^1 -regularity of \bar{p} is not sufficient to prove H^1 -regularity of \bar{u} , because the trace $\bar{p}|_\Gamma \in H^{1/2}(\Gamma)$ appears in the projection formula. Our previous analysis delivers an error estimate of order $\sigma^{1/2}$. Imposing some extra regularity to the data in the elliptic equation, this non-optimal order can be improved.

Lemma 1.2 *Assume that the coefficients a_{ij} of A are Lipschitz functions, that (A2) is satisfied and $\Omega \subset \mathbb{R}^N$ is bounded with boundary of class $C^{1,1}$. Then the solution \bar{p} of (1.11) is in $H^2(\Omega)$. If Ω is a bounded Lipschitz domain and $A = -\Delta$, then \bar{p} belongs to $H^{3/2}(\Omega)$.*

Proof: The result for a $C^{1,1}$ -boundary follows from [8], Thm. 2.4.2.6. If Ω is a bounded Lipschitz domain, then we consider the boundary condition in (1.11) as Neumann condition with right hand side $-\beta \bar{p}$. The result follows from a theorem by [9]. ■

Consequently, the trace $\bar{p}|_\Gamma$ belongs to $H^{3/2}(\Gamma) \subset H^1(\Gamma)$. From the continuity of the projection mapping (1.12) in $H^1(\Gamma)$ we conclude that

$\bar{u} \in H^1(\Gamma)$, if the assumptions of Lemma 1.2 are satisfied. Adapting the proof of Theorem 1.2, we find $\lambda \|\bar{u}_\sigma - \bar{u}\|_{L^2(\Gamma)} \leq (b \bar{p}(y_\sigma), \Pi_\sigma \bar{u} - \bar{u})_{L^2(\Gamma)}$, where $\bar{p}_\sigma = p(\bar{y}_\sigma)$ solves the adjoint equation (1.11) with \bar{y}_σ substituted for \bar{y} . Continuing the proof of Theorem 1.2, we obtain

Theorem 1.3 *Let the assumptions of Lemma 1.2 be satisfied and \bar{u} and \bar{u}_σ be the optimal controls of (P^2) and (P_σ^2) , respectively. Then*

$$\|\bar{u} - \bar{u}_\sigma\|_{L^2(\Gamma)} \leq c_2 \sigma.$$

holds with a constant c_2 that does not depend on σ and \bar{u}_σ .

Boundary observation. We consider (P^2) with the functional

$$J(y, u) = \|y - y_\Gamma\|_{L^2(\Gamma)}^2 + \lambda \|u\|_{L^2(\Gamma)}^2.$$

Then the control-to-observation mapping is $S u = y|_\Gamma$, and we have $S : L^2(\Gamma) \rightarrow H^{1/2}(\Gamma)$. However, we consider S as a mapping in $L^2(\Gamma)$, thus also $S^* : L^2(\Gamma) \rightarrow L^2(\Gamma)$. The adjoint state \bar{p} solves

$$A^* \bar{p} = 0, \quad \partial_{\nu_{A^*}} \bar{p} + \beta \bar{p} = \bar{y}|_\Gamma - y_\Gamma. \quad (1.13)$$

If $y_\Gamma \in L^2(\Gamma)$, we obtain only $\bar{y}|_\Gamma - y_\Gamma \in L^2(\Gamma)$. Even in the case of a regular boundary this would imply at best $\bar{p} \in H^{3/2-\varepsilon}(\Omega)$, hence $\bar{p}|_\Gamma \in H^{1-\varepsilon}(\Gamma)$ and the projection formula (1.12) would not ensure $\bar{u} \in H^1(\Gamma)$. Therefore, we assume $y_\Gamma \in H^{1/2}(\Gamma)$.

Lemma 1.3 *Assume that $\Omega \subset \mathbb{R}^N$ is bounded with boundary of class $C^{1,1}$, the coefficients a_{ij} are Lipschitz functions, (A2) is satisfied, and $y_\Gamma \in H^{1/2}(\Gamma)$. Then the solution \bar{p} of (1.13) belongs to $H^2(\Omega)$. If Ω is a bounded Lipschitz domain and $A = -\Delta$, then $\bar{p} \in H^{3/2}(\Omega)$.*

The first part follows directly from [8], Thm. 2.4.2.6. and Thm. 5.1.3.1, the second from [9].

Remark: For our purposes, it is sufficient to have $\bar{p} \in H^{3/2}(\Omega)$, which gives $\bar{p}|_\Gamma \in H^1(\Gamma)$. This regularity follows from [11], if Γ is of class C^∞ and $y_\Gamma \in H^s(\Gamma)$ for some $s > 0$.

Knowing $\bar{u} \in H^1(\Gamma)$, the error estimate can be proved as in Theorem 1.2. Therefore, the estimate of Theorem 1.2 remains true for boundary observation, if the assumptions of Lemma 1.3 are satisfied.

2. Boundary control by piecewise linear control functions

Regular domains in \mathbb{R}^2 and boundary observation. In the case of polygonal or polyhedral domains, discontinuous functions such as step

functions will not provide the H^2 -regularity needed to perform the error estimates for the application of FEM. Therefore, we also consider piecewise linear controls. We begin with a regular boundary and investigate the most delicate problem – boundary control and boundary observation. The associated case of distributed observation is covered for step functions by Lemma 1.2. Therefore, we consider the problems (P^2) and (P_σ^2) , with boundary observation and another definition for U_σ^{ad} . Let Γ be represented by a closed parametrized curve $x = x(s)$ with arc length $s \in [0, L]$, where L is the length of Γ . We subdivide $[0, L]$ by a partition of mesh size σ , $0 = s_0 < s_1 < \dots < s_m = L$ and define $x_i = x(s_i)$, $i = 0, \dots, m$. Notice that $x_0 = x_m$. Moreover, we put $x_{m+i} = x_i$. Let us identify $\bar{\Gamma}_i$ with the (curved) interval $[x_{i-1}, x_i]$. For the controls u on Γ we write $u = u(x)$ or $u = u(x(s)) =: u(s)$. We work with the set of piecewise linear controls

$$U_\sigma^{lin} = \{u \in U^{ad} \mid u \in C(\Gamma) \text{ and } u \in \mathcal{P}^1(\bar{\Gamma}_j) \forall j = 1, \dots, m\},$$

where $\mathcal{P}^1(\bar{\Gamma}_j)$ stands for the set of polynomials $u = u(s)$ on $[s_{j-1}, s_j]$ of order ≤ 1 on $\bar{\Gamma}_j$. The other notations are adopted from Section 1.2. The functions of U_σ^{lin} are uniquely determined by their values in the x_i .

The proof of Theorem 1.2 must be slightly changed, since we cannot employ the orthogonality relation (1.9). The best approximation of \bar{u} by piecewise linear functions does possibly not belong to U^{ad} . We use the following piecewise linear function $u_\sigma \in U_\sigma^{ad}$: Assuming \bar{u} as continuous, for all sufficiently small $\sigma > 0$ we define

$$u_\sigma(x_i) = \begin{cases} u_a & \text{if } \min_{[x_{i-1}, x_{i+1}]} \bar{u}(x) = u_a \\ u_b & \text{if } \max_{[x_{i-1}, x_{i+1}]} \bar{u}(x) = u_b \\ \bar{u}(x_i) & \text{else.} \end{cases}$$

The mesh size σ must be small such that $\bar{u} = u_a$ and $\bar{u} = u_b$ cannot happen in the same $\bar{\Gamma}_i$.

Lemma 2.1 *Let \bar{p} be the adjoint state obtained from (1.13). Then*

$$(b\bar{p} + \lambda\bar{u}, v - u_\sigma)_{L^2(\Gamma)} \geq 0 \quad (2.14)$$

holds for all sufficiently small $\sigma > 0$ and all $v \in U^{ad}$.

Proof: We fix i and show $(b\bar{p} + \lambda\bar{u})(v - u_\sigma) \geq 0$ on Γ_i : If $\bar{u}(\hat{x}) = u_a$ holds for an $\hat{x} \in [x_{i-1}, x_{i+1}] =: I$, then $\bar{u}(x) < u_b$ on Γ_i follows by continuity for small σ . Then the variational inequality for \bar{u} can only hold, if $b\bar{p} + \lambda\bar{u} \geq 0$ on Γ_i , hence $(b\bar{p} + \lambda\bar{u})(v - u_\sigma) \geq 0$, as $u_\sigma = u_a$ on I . The case $\bar{u}(\hat{x}) = u_b$ is discussed analogously. If $u_a < \bar{u} < u_b$ everywhere in I , then $b\bar{p} + \lambda\bar{u} = 0$ on Γ_i , and the desired inequality is trivial. ■

Now we use (2.14), $b\bar{p} = S^*(S\bar{u} - y_\Gamma)$, write down the variational inequality from the optimality condition for \bar{u}_σ ,

$$\begin{aligned}(S^*(S\bar{u} - y_\Gamma) + \lambda\bar{u}, \bar{u}_\sigma - u_\sigma)_{L^2(\Gamma)} &\geq 0 \\ (S^*(S\bar{u}_\sigma - y_\Gamma) + \lambda\bar{u}_\sigma, u_\sigma - \bar{u}_\sigma)_{L^2(\Gamma)} &\geq 0,\end{aligned}$$

add both relations and obtain after some simple calculations

$$\begin{aligned}\|S(u_\sigma - \bar{u}_\sigma)\|_{L^2(\Gamma)}^2 + \lambda\|u_\sigma - \bar{u}_\sigma\|_{L^2(\Gamma)}^2 &\leq \\ \leq (S^*S(\bar{u} - u_\sigma) + \lambda(\bar{u} - u_\sigma), \bar{u}_\sigma - u_\sigma)_{L^2(\Gamma)}.\end{aligned}$$

Therefore, $\|u_\sigma - \bar{u}_\sigma\|_{L^2(\Gamma)}^2 \leq c\|\bar{u} - u_\sigma\|_{L^2(\Gamma)}\|u_\sigma - \bar{u}_\sigma\|_{L^2(\Gamma)}$ holds, and by the triangle inequality we finally arrive at

$$\|\bar{u} - \bar{u}_\sigma\|_{L^2(\Gamma)} \leq c\|\bar{u} - u_\sigma\|_{L^2(\Gamma)}. \quad (2.15)$$

This is the key relation to prove the following error estimate.

Theorem 2.1 *Assume that the coefficients a_{ij} of A are Lipschitz functions, that (A2) is satisfied and $\Omega \subset \mathbb{R}^2$ is bounded with boundary of class $C^{1,1}$. Let \bar{u} and \bar{u}_σ be the optimal controls of (P^2) and (P_σ^2) , respectively, where U_σ^{lin} is substituted for U_σ^{ad} . Then there is a constant c_L that does not depend on σ and \bar{u}_σ , such that*

$$\|\bar{u} - \bar{u}_\sigma\|_{L^2(\Gamma)} \leq c_L \sigma. \quad (2.16)$$

Proof: We show that a constant c exists such that $\|\bar{u} - u_\sigma\|_{L^2(\Gamma)} \leq c\sigma$. Then the result follows directly from the estimate (2.15). We fix an arbitrary Γ_i and distinct between three cases:

(i) $u_a < \bar{u}(x) < u_b$ everywhere in $[x_{i-2}, x_{i+1}]$: Here, by definition, u_σ coincides on Γ_i with the interpolate of \bar{u} . Therefore,

$$\|\bar{u} - u_\sigma\|_{L^2(\Gamma_i)} = \|\bar{u} - \Pi_\sigma^1 \bar{u}\|_{L^2(\Gamma_i)} \leq c\sigma \|\bar{u}\|_{H^1(\Gamma_i)}.$$

(ii) $\bar{u}(\hat{x}) = u_a$ in some $\hat{x} = x(\hat{s}) \in [x_{i-2}, x_{i+1}]$: If $\hat{x} \in \bar{\Gamma}_i$, then $u_\sigma = u_a$ on $\bar{\Gamma}_i$ and

$$\begin{aligned}\|\bar{u} - u_\sigma\|_{L^2(\Gamma_i)}^2 &= \int_{s_{i-1}}^{s_i} |\bar{u}(s) - \bar{u}(\hat{s})|^2 |x'| ds = \int_{s_{i-1}}^{s_i} \left| \int_s^{\hat{s}} \frac{d}{dt} \bar{u}(x(t)) dt \right|^2 |x'| ds \\ &\leq \sigma \|x'\|_{L^\infty}^2 \int_{s_{i-1}}^{s_i} dt \int_{s_{i-1}}^{s_i} |\nabla \bar{u}(x(t))|^2 |x'(t)| dt \leq c\sigma^2 \|\bar{u}\|_{H^1(\Gamma_i)}^2.\end{aligned}$$

If \bar{u} attains u_a in $[x_{i-2}, x_{i-1}]$ and $[x_i, x_{i+1}]$, then $u_\sigma = u_a$ on $[x_{i-2}, x_{i+1}]$, and we can estimate as above with a certain $\hat{x} \in [x_{i-2}, x_{i+1}] \setminus \Gamma_i$. Here, the integral over $|\nabla \bar{u}|^2$ can be estimated by one over Γ_i and a neighboring Γ_j . We get an estimate by $2c\sigma^2 \|\bar{u}\|_{H^1(\Gamma_i \cup \Gamma_j)}^2$.

If \bar{u} attains u_a only in one of the two neighboring intervals, say in $[x_{i-2}, x_{i-1}]$, then $u_\sigma(x_{i-1}) = u_a = \bar{u}(\hat{x}) = \bar{u}(x(\hat{s}))$, $u_\sigma(x_i) = \bar{u}(x_i)$ and u_σ is affine-linear on Γ_i . We find

$$\begin{aligned} |\bar{u}(x) - u_\sigma(x)| &\leq |\frac{s_i-s}{\sigma}(\bar{u}(s) - u_\sigma(s_{i-1}))| + |\frac{s-s_{i-1}}{\sigma}(\bar{u}(s) - u_\sigma(s_i))| \\ &\leq |\bar{u}(s) - \bar{u}(\hat{s})| + |\bar{u}(s) - \bar{u}(s_i)|. \end{aligned}$$

Proceeding as above, we get an estimate by $3c\sigma^2\|\bar{u}\|_{H^1(\Gamma_i \cup \Gamma_j)}^2$. The same holds, if \bar{u} attains u_a only in $]x_i, x_{i+1}]$. In all the cases the estimate

$$\|\bar{u} - u_\sigma\|_{L^2(\Gamma_i)}^2 \leq 3\sigma^2\|\bar{u}\|_{H^1(\Gamma_i \cup \Gamma_j)}^2$$

is true. Each Γ_i can appear at most twice in this procedure so that, summing up over all i , finally the desired estimate is obtained. ■

2.1. Convex polygonal domains in \mathbb{R}^2 and boundary observation

We only briefly address the case of a polygonal domain Ω . The construction of the last subsection can be applied in almost the same way. However, the regularity properties are more delicate. To perform the error analysis, we need $\bar{u} \in H^1(\Gamma)$ and $\bar{p}|_\Gamma \in H^1(\Gamma)$.

Theorem 2.2 *If $A = -\Delta$, Ω is a bounded convex polygonal domain in \mathbb{R}^2 and $y_\Gamma \in H^{1/2}(\Gamma)$, then the result of Theorem 2.1 remains valid.*

Proof: Since $y_\Gamma \in H^{1/2}(\Gamma)$, the boundary data in the adjoint equation (1.13) belong to $H^{1/2}(\Gamma)$. Therefore, by Theorem 5.1.2.4 of [8], \bar{p} belongs to $H^2(\Omega)$ and its trace is at least in $H^1(\Gamma)$. The projection formula for \bar{u} yields that $\bar{u} \in H^1(\Gamma)$ and this was the only assumption needed to prove Theorem 2.1. ■

Discretization of the elliptic equation. Let us briefly comment on the second step of our analysis – the approximation of the elliptic boundary value problem. Here, we assume that σ , the discretization parameter of controls, is fixed. Behind this is the idea "first discretize the controls, then discretize the state". After σ has been chosen, the mesh size h for the state can be adapted as fine as necessary.

We have studied FEM for the equation under the assumption that Ω is polygonal. Here, h was the mesh size of a regular triangulation of Ω . The admissible sets of discretized controls have a certain smoothness, which is helpful to derive optimal error estimates, namely $U_\sigma^{ad} \subset H^{1/2-\varepsilon}(\Gamma)$ and $U_\sigma^{lin} \subset H^1(\Gamma)$. We found that, for σ fixed, the contribution of the FEM to the error was of order h^2 for (P^1) with distributed observation,

$N = 2, 3$, using step functions. This looks somehow surprising. However, it is not. Notice that we do not approximate the controls here, producing an error of order σ . The controls just *are* discretized. Only the equation is approximated, and this approximation is of order h^2 in $L^2(\Omega)$, because y has optimal regularity $H^2(\Omega)$. For (P^2) with boundary observation and piecewise linear functions, the error had the order $h^{3/2}$. The lower order $h^{3/2}$ comes from the approximation of traces. Combining all results we obtained estimates of the type

$$\|\bar{u} - \bar{u}_{\sigma,h}\|_{L^2} \leq \alpha_1 \sigma + \alpha_2 h^s,$$

where $\bar{u}_{\sigma,h}$ stands for the optimal control of the fully discretized problem, $s = 2$ or $s = 3/2$, and the α_i are independent of σ , h , and $\bar{u}_{\sigma,h}$. This estimate indicates that there is no need to choose the order of discretization for the state larger than for the controls.

Acknowledgement. We are grateful to M. Costabel for drawing our attention to the regularity result by [9].

References

- [1] Arada, N., Casas, E., and Tröltzsch, F. (2002). Error estimates for the numerical approximation of a semilinear elliptic control problem. *Computational Optimization and Applications*, 23:201–229.
- [2] Arnautu, V. and Neittaanmäki, P. (1998). Discretization estimates for an elliptic control problem. *Numer. Funct. Anal. and Optimiz.*, 19:431–464.
- [3] Casas, E. (1992). *Introducción a las ecuaciones en derivadas parciales*. Universidad de Cantabria, Santander.
- [4] Casas, E. and Mateos, M. (2001). Uniform convergence of the FEM. Applications to state constrained control problems. *Comp. Appl. Math.*, *To appear*.
- [5] Ciarlet, P. and Lions, J. (1991). *Handbook of Numerical Analysis, Vol. II, Part I – Finite Element Methods*. North-Holland, Amsterdam.
- [6] Falk, F. (1973). Approximation of a class of optimal control problems with order of convergence estimates. *J. Math. Anal. Appl.*, 44:28–47.
- [7] Geveci, T. (1979). On the approximation of the solution of an optimal control problem governed by an elliptic equation. *R.A.I.R.O. Analyse numérique / Numerical Analysis*, 13:313–328.
- [8] Grisvard, P. (1985). *Elliptic Problems in Nonsmooth Domains*. Pitman, Boston.
- [9] Jerison, D. and Kenig, C. (1981). The Neumann problem on Lipschitz domains. *Bull. Amer. Math. Soc. (N.S.)*, 4:203–207.
- [10] Kinderlehrer, D. and Stampacchia, G. (1980). *An introduction to variational inequalities and their applications*. Academic Press, New York.
- [11] Triebel, H. (1995). *Interpolation theory, function spaces, differential operators*. J. A. Barth Verlag, Heidelberg-Leipzig.

RELAXATION FOR QUASI-LINEAR DIFFERENTIAL INCLUSIONS IN NON SEPARABLE BANACH SPACES

Aurelian Cernea

Faculty of Mathematics, University of Bucharest

Academiei 14, 70109 Bucharest, Romania

acernea@math.math.unibuc.ro

Abstract We consider a Cauchy problem for nonconvex quasi-linear differential inclusions in non separable Banach spaces and we prove that the set of mild solutions of this problem is dense in the set of mild solutions of the convexified problem.

Keywords: quasi-linear inclusions, mild solutions, Lusin measurable multifunctions, relaxation property.

1. Introduction

In this paper we consider quasi-linear differential inclusions of the form

$$x'(t) \in A(t, x(t))x(t) + F(t, x(t)), \quad a.e.([0, T]), \quad x(0) = a, \quad (1.1)$$

where $A(t, w)$ is a linear operator from the real Banach space X in X , $t \in I := [0, T]$, w belongs to an open and nonempty set $D \subset X$ and F is a set-valued map from $I \times X$ to $\mathcal{P}(X)$.

Qualitative properties and structure of the set of solutions of this problem have been studied by many authors ([1], [2], [8], [9], [10], [12] etc). In [9] it is shown that if X is a separable Banach space, the set of mild solutions of the problem (1.1) is dense in the set of mild solutions of the convexified (relaxed) problem:

$$x'(t) \in A(t, x(t))x(t) + \overline{\text{co}}F(t, x(t)), \quad a.e.([0, T]), \quad x(0) = a. \quad (1.2)$$

If the operator A depends neither on t nor on w the quasi-linear inclusions (1.1)-(1.2) reduce to the corresponding semilinear problems

$$x'(t) \in Ax(t) + F(t, x(t)), \quad a.e.([0, T]), \quad x(0) = a, \quad (1.3)$$

$$x'(t) \in Ax(t) + \overline{co}F(t, x(t)), \quad a.e.([0, T]), \quad x(0) = a. \quad (1.4)$$

Recently, De Blasi and Pianigiani ([4]) established a relaxation result for the problems (1.3), (1.4) in an arbitrary, not necessarily separable, Banach space X . Even if the general ideas of proving a relaxation theorem are still present, the approach in [4] has a fundamental difference which consists in the construction of the measurable selections of the multifunction. This construction does not use classical selection theorems as Kuratowsky and Ryll-Nardzewski ([7]) or Bressan and Colombo ([3]).

The aim of this paper is to extend the result in [4] to the more general problem (1.1). We will prove the relaxation property of mild solutions for problem (1.1) in an arbitrary space X . The proof of our main result follows the general ideas in [4] and [9].

2. Preliminaries

Consider X an arbitrary real Banach space with norm $\|.\|$ and let $\mathcal{P}(X)$ be the space of all bounded nonempty subsets of X endowed with the Hausdorff pseudometric

$$d_H(A, B) = \max\{d^*(A, B), d^*(B, A)\}, \quad d^*(A, B) = \sup_{a \in A} d(a, B),$$

where $d(x, A) = \inf_{a \in A} \|x - a\|$, $A \subset X, x \in X$.

Let \mathcal{L} be the σ -algebra of the (Lebesgue) measurable subsets of R and, for $A \in \mathcal{L}$, let $\mu(A)$ be the Lebesgue measure of A .

Let Y be a metric space. An open (resp. closed) ball in Y with center y and radius r is denoted by $B_Y(y, r)$ (resp. $\overline{B}_Y(y, r)$). For any set $A \subset Y$ we denote by $\overline{co}A$ the closed convex hull of A . In what follows $B = B_X(0, 1)$ and $I = [0, 1]$.

A multifunction $F : Y \rightarrow \mathcal{P}(X)$ with closed bounded nonempty values is said to be d_H -continuous at $y_0 \in Y$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that for any $y \in B_Y(y_0, r)$ we have $d_H(F(y), F(y_0)) \leq \epsilon$. F is called d_H -continuous if it is so at each point $y_0 \in Y$.

Let $A \in \mathcal{L}$, with $\mu(A) < \infty$. A multifunction $F : A \rightarrow \mathcal{P}(X)$ with closed bounded nonempty values is said to be *Lusin measurable* if for every $\epsilon > 0$ there exists a compact set $K_\epsilon \subset A$, with $\mu(A \setminus K_\epsilon) < \epsilon$ such that F restricted to K_ϵ is d_H -continuous.

It is clear that if $F, G : A \rightarrow \mathcal{P}(X)$ and $f : A \rightarrow X$ are Lusin measurable then so are F restricted to B ($B \subset A$ measurable), $F + G$ and $t \mapsto d(f(t), F(t))$. Moreover, the uniform limit of a sequence of Lusin measurable multifunctions is also Lusin measurable.

As usual we denote by $L^1(I, X)$ the Banach space of all Bochner integrable functions $x(\cdot) : I \rightarrow X$ endowed with the norm $\|x(\cdot)\|_1 = \int_0^1 \|x(s)\| ds$. In what follows we shall use the following assumptions.

Hypothesis 2.1. a) $F : I \times X \rightarrow \mathcal{P}(X)$ is a set-valued map with closed bounded nonempty values and for any $x \in X$, $F(., x)$ is Lusin measurable on I .

b) There exists an integrable function $k(.) \in L^1(I, R)$ such that

$$d_H(F(t, x), F(t, y)) \leq k(t) \|x - y\| \quad \forall (t, x), (t, y) \in I \times X.$$

c) There exists $q(.) \in L^1(I, R)$ such that for any continuous function $x(.) \in C(I, D)$ and any $t \in I$ we have

$$F(t, x(t)) \subset q(t)B.$$

A family of bounded linear operators $\mathcal{U}(t, s)$ on X , $0 \leq s \leq t \leq 1$ depending on two parameters is said to be an *evolution system* ([11]) if the following conditions are satisfied

- 1) $\mathcal{U}(s, s) = 1$, $\mathcal{U}(t, r)\mathcal{U}(r, s) = \mathcal{U}(t, s)$ $0 \leq s \leq r \leq t \leq T$
- 2) $(t, s) \rightarrow \mathcal{U}(t, s)$ is strongly continuous, i.e.

$$\lim_{t \searrow s} \mathcal{U}(t, s)x = x, \quad \forall x \in X.$$

Hypothesis 2.2. a) X is a real Banach space and $D \subset X$ is an open nonempty set.

b) For any $u \in C(I, D)$ the family of linear operators $\{A(t, u), t \in I\}$ generates a unique strongly continuous evolution system $\mathcal{U}_u(t, s)$, $0 \leq s \leq t \leq T$.

c) If $u \in C(I, D)$, the evolution system $\mathcal{U}_u(t, s)$, $0 \leq s \leq t \leq T$ satisfies:

there exists $M \geq 0$ such that $\|\mathcal{U}_u(t, s)\| \leq M$, $0 \leq s \leq t \leq T$, uniformly in u ;

for any $u, v \in C(I, D)$ and any $w \in D$ we have

$$\|\mathcal{U}_u(t, s)w - \mathcal{U}_v(t, s)w\| \leq M\|w\| \int_s^t \|u(\tau) - v(\tau)\| d\tau.$$

By a *mild solution* of the Cauchy problem (1.1) we mean a function $x(.) : I \rightarrow X$ satisfying the following conditions:

- i) $x(.)$ is continuous on I with $x(0) = a$
- ii) there exists a Lusin measurable function $f(.) : I \rightarrow X$, Bochner integrable such that

$$f(t) \in F(t, x(t)), \quad \forall t \in I,$$

$$x(t) = \mathcal{U}_x(t, 0)a + \int_0^t \mathcal{U}_x(t, s)f(s)ds, \quad \forall t \in I.$$

According to [4] in the above definition the Lusin measurability of $f(\cdot)$ is equivalent to the (strong) measurability of $f(\cdot)$.

We recall some results we shall use in the sequel.

Lemma 2.3. ([4]) *Let $F : I \times X \rightarrow \mathcal{P}(X)$ be a set-valued map with closed bounded nonempty values that satisfies Hypothesis 2.1. Then, for any $x(\cdot) : I \rightarrow X$ continuous, $u(\cdot) : I \rightarrow X$ measurable and $\epsilon > 0$ we have:*

- a) *the multifunction $t \rightarrow F(t, x(t))$ is Lusin measurable on I .*
- b) *the multifunction $G : I \rightarrow \mathcal{P}(X)$ defined by*

$$G(t) := (F(t, x(t)) + \epsilon B) \cap B_X(u(t), d(u(t), F(t, x(t))) + \epsilon)$$

has a Lusin measurable selection $f : I \rightarrow X$.

Lemma 2.4. ([9]) *Suppose that Hypothesis 2.2 is satisfied and that each quasi-linear Cauchy problem*

$$x'_n(t) \in A(t, x_n(t))x_n(t) + f_n(t), \quad a.e.([0, T]), \quad x_n(0) = a,$$

$n \in N$, has a mild solution

$$x_n(t) = \mathcal{U}_{x_n}(t, 0)a + \int_0^t \mathcal{U}_{x_n}(t, s)f_n(s)ds, \quad t \in I.$$

Suppose, also, that there exists $x \in C(I, X)$ and $f \in L^1(I, X)$ such that $x_n \rightarrow x$ in $C(I, X)$, $f_n \rightarrow f$ in $L^1(I, X)$ and the set $\{f\} \cup \{f_n\}_{n \in N}$ is integrably bounded by a function $m \in L^1(I, X)$. Then,

$$x(t) = \mathcal{U}_x(t, 0)a + \int_0^t \mathcal{U}_x(t, s)f(s)ds, \quad \forall t \in I.$$

3. The main result

In order to prove our main result we need the following lemma which is a quasi-linear version of Lemma 4.2 in [4], obtained for linear differential inclusions. The proof can be easily performed through the same arguments employed to establish Lemma 4.2 in [4].

Lemma 3.1. *We assume that Hypotheses 2.1-2.2 are satisfied. Let $a \in X$ and let $y(\cdot) : I \rightarrow X$ be a mild solution of the relaxed problem (1.2). Then, for any $0 < \sigma < 1$ there exists a mild solution $x_0(\cdot) : I \rightarrow X$ of the Cauchy problem*

$$x'(t) \in A(t, x(t))x(t) + F(t, x(t)) + \phi_\sigma(t)B, \quad a.e.([0, T]), \quad x(0) = a, \quad (3.1)$$

where $\phi_\sigma(\cdot) \in L^1(I, [0, \infty))$ with $\int_0^1 \phi_\sigma(t)dt < 2\sigma$, such that

$$\|x_0(t) - y(t)\| < \sigma, \quad \forall t \in I.$$

Our main result states that the set of mild solutions of the problem (1.1) is dense in the set of mild solutions of the convexified (relaxed) problem (1.2).

Theorem 3.1. *We assume that Hypotheses 2.1-2.2 are satisfied. Let $a \in X$ and let $y(\cdot) : I \rightarrow X$ be a mild solution of the convexified problem (1.2). Then, for every $\epsilon > 0$ there exists a mild solution $x(\cdot) : I \rightarrow X$ of the problem (1.1) such that:*

$$\|x(t) - y(t)\| < \epsilon, \quad \forall t \in I.$$

Proof. Let $y(\cdot) : I \rightarrow X$ be an arbitrary mild solution of the Cauchy problem (1.2) and let $0 < \epsilon < 1$. We define

$$L(t) := \int_0^t M(\|a\| + \|q\|_1 + \epsilon + k(s))ds, \quad t \in I.$$

Fix σ such that $0 < \sigma < \frac{\epsilon}{(3M+1)e^{L(1)}}$. Let $\phi_\sigma(\cdot) \in L^1(I, [0, \infty))$ such that $\int_0^1 \phi_\sigma(t)dt < 2\sigma$.

By Lemma 3.1 there exists a mild solution $x_0(\cdot) : I \rightarrow X$ of the problem (3.1) such that:

$$\|x_0(t) - y(t)\| < \sigma, \quad \forall t \in I. \quad (3.2)$$

By definition of mild solution $x_0(\cdot)$ is continuous, $x_0(0) = a$ and there exists a Lusin measurable function $f_0(\cdot) : I \rightarrow X$, Bochner integrable such that

$$f_0(t) \in F(t, x_0(t)) + \phi_\sigma(t)B, \quad t \in I, \quad (3.3)$$

$$x_0(t) = \mathcal{U}_{x_0}(t, 0)a + \int_0^t \mathcal{U}_{x_0}(t, s)f_0(s)ds, \quad t \in I. \quad (3.4)$$

Let $\sigma_n = \frac{\sigma}{2^{n+2}}$ and $p_0(t) := d(f_0(t), F(t, x_0(t))), t \in I$.

Since $x_0(\cdot)$ is continuous, by Lemma 2.3 there exists a Lusin measurable function $f_1(\cdot) : I \rightarrow X$ satisfying, for $t \in I$,

$$f_1(t) \in (F(t, x_0(t)) + \epsilon_1 B) \cap B_X(f_0(t), d(f_0(t), F(t, x_0(t))) + \sigma_1)$$

Hence $f_1(\cdot)$ is also Bochner integrable on I . Define $x_1(\cdot) : I \rightarrow X$ by

$$x_1(t) = \mathcal{U}_{x_0}(t, 0)a + \int_0^t \mathcal{U}_{x_0}(t, s)f_1(s)ds, \quad \forall t \in I.$$

By reccurrence, we construct a sequence $\{x_n\}_n$ of continuous functions $x_n : I \rightarrow X, n \geq 2$ given by

$$x_n(t) = \mathcal{U}_{x_{n-1}}(t, 0)a + \int_0^t \mathcal{U}_{x_{n-1}}(t, s)f_n(s)ds, \quad t \in I, \quad (3.5)$$

with $f_n(\cdot) : I \rightarrow X$ a Lusin measurable function satisfying, for $t \in I$,

$$f_n(t) \in (F(t, x_{n-1}(t)) + \sigma_n B) \cap B_X(f_{n-1}(t), d(f_{n-1}(t), F(t, x_{n-1}(t))) + \sigma_n). \quad (3.6)$$

From (3.6), for $n \geq 2$, we obtain

$$\begin{aligned} \|f_n(t) - f_{n-1}(t)\| &\leq d(f_{n-1}(t), F(t, x_{n-1}(t))) + \sigma_n \leq \\ &\leq d(f_{n-1}(t), F(t, x_{n-2}(t))) + d_H(F(t, x_{n-2}(t)), F(t, x_{n-1}(t))) + \sigma_n \\ &\leq \sigma_{n-1} + k(t)\|x_{n-1}(t) - x_{n-2}(t)\| + \sigma_n. \end{aligned}$$

Since $\sigma_{n-1} + \sigma_n < \sigma_{n-2}$ we deduce, for $n \geq 2$, that

$$\|f_n(t) - f_{n-1}(t)\| \leq \sigma_{n-2} + k(t)\|x_{n-1}(t) - x_{n-2}(t)\|. \quad (3.7)$$

Define $r(t) := M \int_0^t (p_0(s) + \sigma) ds$, $t \in I$.

One has

$$\begin{aligned} \|x_1(t) - x_0(t)\| &\leq \int_0^t \|\mathcal{U}_{x_0}(t, s)f_1(s) - \mathcal{U}_{x_0}(t, s)f_0(s)\| ds \leq \\ &\leq M \int_0^t (p_0(s) + \sigma) ds = r(t) \end{aligned}$$

Clearly, by (3.3) $p_0(t) \leq \phi_\sigma(t)$, $\forall t \in I$, hence $\int_0^1 p_0(t) dt \leq \int_0^1 \phi_\sigma(t) dt < 2\sigma$. Thus, $r(1) = M \int_0^1 (p_0(s) + \sigma) ds < 3M\sigma$.

By recurrence, we shall prove that, for $n \geq 1$, one has

$$\|x_n(t) - x_{n-1}(t)\| \leq r(t) \frac{(L(t))^{n-1}}{(n-1)!} \quad \forall t \in I. \quad (3.8)$$

For $n = 1$ the inequality is already proved.

Assuming that (3.8) is valid for n , we show that (3.8) holds for $n+1$. Using (3.5)-(3.8) one has

$$\begin{aligned} \|x_{n+1}(t) - x_n(t)\| &\leq \|\mathcal{U}_{x_n}(t, 0)a - \mathcal{U}_{x_{n-1}}(t, 0)a\| + \\ &\quad \int_0^t \|\mathcal{U}_{x_n}(t, s)f_{n+1}(s) - \mathcal{U}_{x_{n-1}}(t, s)f_n(s)\| ds \leq \\ &\leq M\|a\| \int_0^t \|x_n(s) - x_{n-1}(s)\| ds + \int_0^t \|\mathcal{U}_{x_n}(t, s)f_{n+1}(s) - \mathcal{U}_{x_n}(t, s)f_n(s)\| ds + \\ &\quad + \int_0^t \|\mathcal{U}_{x_n}(t, s)f_n(s) - \mathcal{U}_{x_{n-1}}(t, s)f_n(s)\| ds \leq \\ &\leq M\|a\| \int_0^t \|x_n(s) - x_{n-1}(s)\| ds + M \int_0^t \|f_{n+1}(s) - f_n(s)\| ds + \end{aligned}$$

$$\begin{aligned}
& + \int_0^t M ||f_n(\tau)|| (\int_\tau^t ||x_n(s) - x_{n-1}(s)|| ds) d\tau \leq \\
\leq & M ||a|| \int_0^t ||x_n(s) - x_{n-1}(s)|| ds + M \int_0^t (\sigma_{n-1} + k(s)) ||x_n(s) - x_{n-1}(s)|| ds + \\
& + \int_0^t M (||q||_1 + \sigma_n) ||x_n(s) - x_{n-1}(s)|| ds \leq \int_0^t [M (||a|| + ||q||_1 + \sigma_{n-2}) + \\
& + M k(s)] ||x_n(s) - x_{n-1}(s)|| ds \leq \int_0^t L'(s) ||x_n(s) - x_{n-1}(s)|| ds \leq \\
\leq & \int_0^t L'(s) r(s) \frac{(L(s))^{n-1}}{(n-1)!} ds < r(t) \frac{(L(t))^n}{n!}.
\end{aligned}$$

From (3.8) we obtain

$$\begin{aligned}
||x_n(t) - x_0(t)|| & \leq \sum_{k \geq 1} ||x_k(t) - x_{k-1}(t)|| \leq \\
& \leq r(t) \sum_{k=1}^{n-1} \frac{(L(t))^k}{k!} < r(1) e^{L(1)}.
\end{aligned}$$

Therefore

$$||x_n(t) - x_0(t)|| \leq 3M e^{L(1)} \sigma, \quad t \in I. \quad (3.9)$$

On the other hand, from (3.4) it follows that

$$||x_{n+1}(t) - x_n(t)|| \leq r(1) \frac{(L(1))^n}{n!} \quad t \in I. \quad (3.10)$$

From (3.10) it follows that the sequence $\{x_n\}_n$ converges uniformly on I to a continuous function, $x(\cdot) : I \rightarrow X$.

In view of (3.7) we have

$$||f_{n+1}(t) - f_n(t)|| \leq \sigma_{n-1} + k(t) r(1) \frac{(L(1))^n}{n!}, \quad t \in I, \quad (3.6)$$

which implies that the sequence $\{f_n\}_n$ converges to a Lusin measurable function $f(\cdot) : I \rightarrow X$. From (3.6) it follows that

$$||f_n(t)|| \leq ||q(t)|| + 1 \quad \forall t \in I, n \in N,$$

hence f is Bochner integrable on I .

Letting $n \rightarrow \infty$ and using Lemma 2.4 we conclude that $x(\cdot)$ is a mild solution of the Cauchy problem (1.1).

From (3.9) we infer that

$$||x(t) - x_0(t)|| \leq 3M e^{L(1)} \sigma, \quad t \in I. \quad (3.9)$$

Finally, from the last estimation, (3.2) and the choice of σ we deduce

$$\begin{aligned} ||x(t) - y(t)|| &\leq ||x(t) - x_0(t)|| + ||x_0(t) - y(t)|| \leq \\ &\leq (3M + 1)e^{L(1)}\sigma < \epsilon, t \in I \end{aligned}$$

and the proof is complete.

Remark 3.3. When the operator A depends neither on t nor on w problem (1.1) reduces to problem (1.3) and Theorem 3.1 yields known results, namely Theorem 4.1 in [4].

References

- [1] Ahmed N. U. and Xiang X.: Optimal control of infinite dimensional uncertain systems, *J. Optim. Theory Appl.* **80** (1994), 261–272.
- [2] Anguraj A. and Balachandran K.: Existence of solutions of nonlinear differential inclusions, *Mem. Fac. Sci. Kochi Univ.* **13** (1992), 61–66.
- [3] Bressan A. and Colombo G.: Extensions and selections of maps with decomposable values, *Studia Math.* **90** (1988), 69–86.
- [4] De Blasi F. S. and Pianigiani G.: Evolution inclusions in non separable Banach spaces, *Comment. Math. Univ. Carolinae* **40** (1999), 227–250.
- [5] Filippov A. F.: Classical solutions of differential equations with multi-valued right hand side, *SIAM J. Control* **5** (1967), 609–621.
- [6] Frankowska H.: A priori estimates for operational differential inclusions, *J. Diff. Eqs.* **84** (1990), 100–128.
- [7] Kuratowski K. and Ryll-Nardzewski C.: A general theorem on selectors, *Bull. Acad. Pol. Sci. Math. Astron. Phys.* **13** (1965), 397–403.
- [8] Muresan M.: On a boundary value problem for quasi-linear differential inclusions of evolutions, *Collect Math.* **45** (1994), 165–175.
- [9] Muresan M.: Qualitative properties of solutions to quasi-linear inclusions II, *Pure Math. Appl.* **5** (1994), 331–353.
- [10] Papageorgiu N. S.: On multivalued evolution equations and differential inclusions in Banach spaces, *Comment. Math. Univ. St. Pauli* **36** (1986), 21–39.
- [11] Pazy A.: *Semigroups of linear operators and applications to partial differential equations*, Springer, Berlin, 1983.
- [12] Sanekata N.: Abstract quasi-linear equations of evolutions in nonreflexive Banach spaces, *Hiroshima Math. J.* **19** (1989), 109–139.

DETERMINING FUNCTIONALS FOR A CLASS OF SECOND ORDER IN TIME EVOLUTION EQUATIONS WITH APPLICATIONS TO VON KARMAN EQUATIONS

Igor Chueshov*

Dept. of Mathematics and Mechanics

Kharkov University

Svobody Square 4

Kharkov 61077, Ukraine

chueshov@ilt.kharkov.ua

Irena Lasiecka

Dept. of Mathematics

University of Virginia

Charlottesville VA 22904, USA

il2v@weyl.math.virginia.edu

Abstract In this paper we present a general approach to construction of determining functionals for second order in time evolution equations with nonlinear damping. As an example we consider von Karman evolution equations which describe nonlinear oscillations of an elastic plate.

Keywords: Long-time dynamics, determining parameters, von Karman equations.

Introduction

The question of the number of parameters that are necessary for the description of the long-time behaviour of solutions to nonlinear partial differential equations was first discussed by Foias and Prodi [7] and Ladyzhenskaya [11] for the 2D Navier-Stokes equations. They proved that

*Research partially supported by INTAS Grant 2000/899

the asymptotic behaviour of the solutions is completely determined by dynamics of the first N Fourier modes, if N is sufficiently large. Later a general approach to the problem of the existence of a finite number of determining functionals (parameters) for dissipative evolution PDE has been developed (see [3] and [4, Chap.5] for a survey).

In this paper we present an approach to construction of determining functionals for second order in time evolution equations with *nonlinear* damping. Nonlinear dissipation does not allow to apply general methods developed in [3, 4]. As an example von Karman evolution equations which describe nonlinear plate oscillations is considered.

As in [2, 3, 4] we involve the concept of the completeness defect for a description of sets of determining functionals. Assume that X and Y are Banach spaces and X continuously and densely embedded into Y . Let $\mathcal{L} = \{l_j : j = 1, \dots, N\}$ be a finite set of linearly independent continuous functionals on X . We define the completeness defect $\epsilon_{\mathcal{L}}(X, Y)$ of the set \mathcal{L} with respect to the pair of the spaces X and Y by the formula

$$\epsilon_{\mathcal{L}}(X, Y) = \sup\{\|w\|_Y : w \in X, l_j(w) = 0, l_j \in \mathcal{L}, \|w\|_X \leq 1\}.$$

The value $\epsilon_{\mathcal{L}}$ is proved to be very useful for characterization of sets of determining functionals (see, e.g., [3, 4] and the references therein). One can show that the completeness defect $\epsilon_{\mathcal{L}}(X, Y)$ is the best possible global error of approximation in Y of elements $u \in X$ by elements of the form $u_{\mathcal{L}} = \sum_{j=1}^N l_j(u)\phi_j$, where $\{\phi_j : j = 1, \dots, N\}$ is an arbitrary set in X . The smallness of $\epsilon_{\mathcal{L}}(X, Y)$ is the main condition (see the results presented below) that guarantee the property of a set of functionals to be asymptotically determining. The so-called modes, nodes and local volume averages (the description of these functionals can be found in [3], for instance) are the main examples of sets of functionals with a small completeness defect. For further discussions and for other properties of the completeness defect we refer to [3, 4]. Here we point out the following estimate

$$\|u\|_Y \leq C_{\mathcal{L}} \cdot \max_{j=1, \dots, N} |l_j(u)| + \epsilon_{\mathcal{L}}(X, Y) \cdot \|u\|_X, \quad u \in X, \quad (1)$$

where $C_{\mathcal{L}} > 0$ is a constant depending on \mathcal{L} . Below we also need the following assertion (see [3]).

Proposition 1. *Let Z be a reflexive Banach spaces such that $X \subset Z \subset Y$ and all the embeddings are continuous and dense. Assume that the inequality*

$$\|u\|_Z \leq a_{\theta} \|u\|_Y^{\theta} \|u\|_X^{1-\theta}, \quad u \in X,$$

is valid with some constants $a_\theta > 0$ and $0 < \theta < 1$. Then for any set \mathcal{L} of the linear functionals on X the estimate $\epsilon_{\mathcal{L}}(X, Z) \leq a_\theta [\epsilon_{\mathcal{L}}(X, Y)]^\theta$ holds.

Abstract model

We consider the following second-order abstract equation:

$$\begin{cases} Mu_{tt}(t) + \mathcal{A}u(t) + Du_t(t) = F(u(t)), \\ u|_{t=0} = u_0, \quad u_t|_{t=0} = u_1, \end{cases} \quad (2)$$

under the following set of assumptions:

Assumption 1 (A1) \mathcal{A} is a closed, linear positive selfadjoint operator acting on a Hilbert space \mathcal{H} with $\mathcal{D}(\mathcal{A}) \subset \mathcal{H}$. We shall denote by $|\cdot|$ and $\|\cdot\|$ the norm of \mathcal{H} and $\mathcal{D}(\mathcal{A}^{\frac{1}{2}})$, respectively; (\cdot, \cdot) will denote a scalar product in \mathcal{H} . We shall use the same symbol to denote the duality pairing between $\mathcal{D}(\mathcal{A}^{\frac{1}{2}})$ and $\mathcal{D}(\mathcal{A}^{\frac{1}{2}})'$.

- (A2) Let V be another Hilbert space such that $\mathcal{D}(\mathcal{A}^{\frac{1}{2}}) \subset V \subset \mathcal{H} \subset V' \subset \mathcal{D}(\mathcal{A}^{\frac{1}{2}})'$, all injections being continuous and dense, $M \in L(V, V')$, the bilinear form (Mu, v) is symmetric and $(Mu, u) \geq \alpha_0 |u|_V^2$, where $\alpha_0 > 0$ and (\cdot, \cdot) is understood as a duality pairing between V and V' . Hence, $M^{-1} \in L(V', V)$. Setting $\bar{M} = M|_{\mathcal{H}}$ with $\mathcal{D}(\bar{M}) = \{u \in V; Mu \in \mathcal{H}\}$ we have $D(\bar{M}^{\frac{1}{2}}) = V$.
- (A3) The operator $D : \mathcal{D}(\mathcal{A}^{1/2}) \rightarrow [\mathcal{D}(\mathcal{A}^{1/2})]'$ is monotone and hemicontinuous with $D(0) = 0$ and $(Du - Dv, u - v) \geq 0$ for $u, v \in \mathcal{D}(\mathcal{A}^{1/2})$.
- (A4) The nonlinear operator $F : \mathcal{D}(\mathcal{A}^{\frac{1}{2}}) \rightarrow V'$ is locally Lipschitz, i.e.:

$$|F(u_1) - F(u_2)|_{V'} \leq L(K) \|u_1 - u_2\|, \quad \forall \|u_i\| \leq K.$$

We also assume that F has the form $F(u) = -\Pi'(u)$, i.e. $F(u)$ is the Frechet derivative of C^1 -functional $\Psi(u) = -\Pi(u)$ on $\mathcal{D}(\mathcal{A}^{\frac{1}{2}})$, where $\Pi(u)$ is bounded on bounded sets from $\mathcal{D}(\mathcal{A}^{1/2})$ and the function $\alpha(\mathcal{A}u, u) + \Pi(u)$ is bounded from below on $\mathcal{D}(\mathcal{A}^{1/2})$ for some $0 \leq \alpha < 1/2$.

It can be shown (see, e.g., [13] and [6, Chap.2]) that under Assumption 1 there exists a global solution $u(t)$ to (2) from $C(\mathbf{R}_+, \mathcal{D}(\mathcal{A}^{\frac{1}{2}})) \cap C^1(\mathbf{R}_+, V)$ and the following energy relation

$$E(u(t), \dot{u}(t)) + \int_0^t (D\dot{u}(\tau), \dot{u}(\tau)) d\tau = E(u_0, u_1) \quad (3)$$

holds, where $E(u_0, u_1) = \frac{1}{2} ((Mu_1, u_1) + (\mathcal{A}u_0, u_0)) + \Pi(u_0)$.

Below we also need the following hypotheses, which are responsible for long time behaviour of solutions to (2).

Assumption 2 (A5) *There exists a continuous, increasing, concave function $H : \mathbf{R}_+ \mapsto \mathbf{R}_+$, $H(0) = 0$, such that*

$$(Mv, v) \leq H((Dv, v)), \quad (4)$$

We also assume that

$$(Dv, u) \leq C(|\mathcal{A}^{1/2}u|) \cdot (Dv, v) + c_2|\mathcal{A}^{1/2-\eta}u|^2, \quad (5)$$

for any $u, v \in \mathcal{D}(\mathcal{A}^{1/2})$, where $C(r)$ is non-decreasing function of $r > 0$, c_2 is a positive constant and $\eta \in (0, 1/2]$.

(A6) *There exist constants $b_0 > 0$ and $\eta \in (0, 1/2]$ such that*

$$(F(w + u) - F(w + z \cdot u), u) \leq (1 - z)b_0(\mathcal{A}^{1-2\eta}u, u), \quad (6)$$

for any $z \in [0, 1]$, $u \in \mathcal{D}(\mathcal{A}^{1/2})$ and w from the set $\mathcal{N} = \{u \in \mathcal{D}(\mathcal{A}^{1/2}) : \mathcal{A}u = F(u)\}$ of stationary solutions.

Let $H_T(s) \equiv 3H(s/T)$, where $T > 0$. Since H_T is increasing, $cI + H_T$ is invertible for every $c > 0$. Therefore the function $p(s) \equiv (cI + H_T)^{-1}(s)$ is positive, continuous and strictly increasing with $p(0) = 0$. Finally we set $q(s) \equiv s - (I + p)^{-1}(s)$ for $s \geq 0$. It is clear that q is strictly increasing, positive and zero at the origin. With function q we associate the nonlinear differential equation:

$$\frac{d}{dt}S(t) + q(S(t)) = f(t), \quad t > 0; \quad S(0) = S_0 \in \mathbf{R}. \quad (7)$$

Since q is monotone increasing, for any $f \in L_1(\mathbf{R}_+)$ there exists unique global solution $S \in C(\mathbf{R}_+)$. Moreover, if $f = 0$ then $S(t) \rightarrow 0$ when $t \rightarrow \infty$. We refer to [10, 13] for details.

Determining functionals

Our main result is the following assertion.

Theorem 3 *Let $u(t)$ be a solution to problem (2) and $w \in \mathcal{D}(\mathcal{A}^{1/2})$ be a solution to the stationary problem $\mathcal{A}u = F(u)$. Assume that $\mathcal{L} = \{l_j : j = 1, \dots, N\}$ is a set of functionals on $\mathcal{D}(\mathcal{A}^{1/2})$ and $\epsilon_{\mathcal{L}} \equiv \epsilon_{\mathcal{L}}(\mathcal{D}(\mathcal{A}^{1/2}), \mathcal{H})$ is the corresponding completeness defect. Assume that Assumption 1 and Assumption 2 hold.*

Part I. We assume in addition that (A.5) in Assumption 2 is satisfied with $H(s) = c_1 s$, $c_1 > 0$. Then the condition

$$\lim_{t \rightarrow +\infty} l_j(u(t)) = l_j(w) \quad \text{for all } j = 1, \dots, N,$$

implies that

$$\lim_{t \rightarrow \infty} (|M^{1/2} u_t(t)|^2 + \|u(t) - w\|^2) = 0 \quad (8)$$

provided $\epsilon_{\mathcal{L}}^{4\eta}(b_0 + 2c_2) < 1$, where η , b_0 , and c_2 are the constants from (5) and (6).

Part II. Assume that there exists a positive function $h(t)$ on \mathbf{R}_+ with the properties: (a) $h(t)$ is decreasing and $\lim_{t \rightarrow \infty} h(t) = 0$; (b) for any $T > 0$ there exists $c_T > 0$ such that $h(t) \leq c_T h(t+T)$ and (c) $\max_j |l_j(u(t) - w)|^2 \leq h(t)$ for $t > 0$. Then there exists $T \geq 2$ such that

$$|M^{1/2} u_t(t)|^2 + \|u(t) - w\|^2 \leq C \cdot \left(S \left(\frac{t}{T} - 1 \right) + h(t) \right)$$

for all $t > T$, where $S(t)$ satisfies the nonlinear ODE (7) (with parameter c defining q depending on the values of constants assumed in Assumption 1 and Assumption 2) and with $f(t) \equiv p(Ch(Tt))$ and $S(0)$ depending on u_0 , u_1 and w . Here C is a constant.

Remark 4 The first part of Theorem 3 refers to the situation of strong dissipation when nonlinear damping Du_t leads to exponential decay rates for the unforced problems. In this case, the mere convergence to zero of $l_j(u(t) - w)$ when $t \rightarrow \infty$ guarantees the asymptotic convergence of $u(t)$ to an equilibrium.

In the case of weaker dissipation $D(u_t)$ (e.g., $H(s)$ is sublinear at the origin) and additional information available on the decay rates of functionals $l_j(u(t))$ to $l_j(w)$, the second part of Theorem 3 provides decay rates for the convergence of solution $u(t)$ to the equilibrium. These rates are described by solutions of a nonlinear ODE (7).

Proof. We rely here on some ideas developed in [5] for wave equation with nonlinear dissipation. Let $v(t) = u(t) - w$. Then for $v(t)$ we have the following equation

$$Mv_{tt} + D(v_t) + \mathcal{A}v = F(w + v(t)) - F(w), \quad t > 0. \quad (9)$$

Multiplying equation (9) in \mathcal{H} by v_t we obtain:

$$\frac{1}{2} \cdot \frac{d}{dt} (|M^{1/2} v_t(t)|^2 + |\mathcal{A}^{1/2} v(t)|^2) + (D(v_t), v_t) = (F(u) - F(w), v_t). \quad (10)$$

It is not difficult to see that

$$(F(u) - F(w), v_t) = (F(u), u_t) - (F(w), v_t) = -\frac{d}{dt}\Phi(v(t)),$$

where

$$\Phi(v) = \Pi(u) - \Pi(w) + (F(w), v) \equiv - \int_0^1 (F(w + zv) - F(w), v) dz.$$

Consequently from (10) we obtain the equality:

$$\frac{d}{dt}\tilde{E}(t) + (D(v_t), v_t) = 0, \quad (11)$$

where

$$\tilde{E}(t) = \frac{1}{2} \left(|M^{1/2}v_t(t)|^2 + |\mathcal{A}^{1/2}v(t)|^2 \right) + \Phi(v(t)).$$

It follows from assumption (A6) that

$$\Phi(v) \geq -\frac{b_0}{2}|\mathcal{A}^{1/2-\eta}v|^2. \quad (12)$$

Since $|\mathcal{A}^{1/2-\eta}v| \leq |\mathcal{A}^{1/2}v|^{1-2\eta} \cdot |v|^{2\eta}$, $0 \leq \eta \leq 1/2$, using (1) with $X = \mathcal{D}(\mathcal{A}^{1/2})$ and $Y = \mathcal{H}$ and Proposition 1 we obtain that

$$|\mathcal{A}^{1/2-\eta}v|^2 \leq (1 + \delta)\epsilon_{\mathcal{L}}^{4\eta}|\mathcal{A}^{1/2}v|^2 + C_{\mathcal{L},\delta} \max_{j=1,\dots,N} |l_j(u)|^2, \quad (13)$$

for each $\delta > 0$. From (12) and (13) we get the following estimate.

Lemma 5

$$\tilde{E}(t) \geq \frac{1}{2}|M^{1/2}v_t|^2 + \left(\frac{1}{2} - \frac{b_0}{2}\epsilon_{\mathcal{L}}^{4\eta}(1+\delta) \right) |\mathcal{A}^{1/2}v|^2 - C_{\mathcal{L},\delta} \max_j |l_j(v)|^2. \quad (14)$$

Moreover, $\tilde{E}(t) \geq 0$ for all $t \geq 0$ provided $\epsilon_{\mathcal{L}}^{4\eta}b_0 < 1$.

Proof. The inequality in (14) is a consequence of inequalities (12) and (13). In order to prove positivity of \tilde{E} we note that it follows from (11) and assumption (A3) that the function $\tilde{E}(t)$ is monotony decreasing. Inequality (14) implies that if $\epsilon_{\mathcal{L}}^{4\eta} < b_0^{-1}$ and $\lim_{t \rightarrow +\infty} l_j(v(t)) = 0$ then $\lim_{t \rightarrow +\infty} \tilde{E}(t) \geq 0$. Thus $\tilde{E}(t) \geq 0$, $t > 0$, as desired. ■

The following estimate is critical for the asymptotic behaviour.

Lemma 6 *Let $T > T_0$, where $T_0 > 0$ is sufficiently large. Then*

$$2p(\tilde{E}(mT)/2) + \tilde{E}(mT) \leq \tilde{E}((m-1)T) + p(\mathcal{N}_{\mathcal{L}}(m, T))$$

for $m = 1, 2 \dots$, where

$$\begin{aligned}\mathcal{N}_{\mathcal{L}}(m, T) &= C_{\mathcal{L}} \left[\max_j |l_j(v(mT))|^2 + \max_j |l_j(v((m-1)T))|^2 \right. \\ &\quad \left. + \int_{(m-1)T}^{mT} \max_j |l_j(v(s))|^2 ds \right].\end{aligned}$$

In particular, when $H(s) = c_1 s$, then we have

$$\tilde{E}(mT) \leq \gamma \tilde{E}((m-1)T) + \mathcal{N}_{\mathcal{L}}(m, T), \quad m = 1, 2 \dots, \quad 0 < \gamma < 1.$$

Proof. Multiplying equation (9) by v we find

$$\frac{d}{dt} (Mv_t, v) = |M^{1/2}v_t|^2 - |A^{1/2}v|^2 + (F(v+w) - F(w), v) - (D(v_t), v).$$

Since $-\frac{1}{2}|A^{1/2}v|^2 = -\tilde{E}(t) + \frac{1}{2}|M^{1/2}v_t|^2 + \Phi(v)$, we have

$$\begin{aligned}\frac{d}{dt} (Mv_t, v) &= \frac{3}{2}|M^{1/2}v_t|^2 - \frac{1}{2}|A^{1/2}v|^2 \\ &\quad + \{\Phi(v) + (F(w+v) - F(w), v)\} - (D(v_t), v) - \tilde{E}(t).\end{aligned}$$

By (A6) we have

$$\begin{aligned}\Phi(v) &+ (F(w+v) - F(w), v) \\ &= \int_0^1 (F(w+v) - F(w+zv)), v) dz \leq \frac{b_0}{2} |A^{1/2-\eta}v|^2.\end{aligned}$$

Therefore using (13) we obtain the following inequality

$$\begin{aligned}\frac{d}{dt} (Mv_t, v) &\leq -\tilde{E}(t) - \frac{1}{2} (1 - \epsilon_{\mathcal{L}}^{4\eta} (1 + \delta) b_0) |A^{1/2}v|^2 \\ &\quad + C_{\mathcal{L}, \delta} \max_j |l_j(v)|^2 + \frac{3}{2} |M^{1/2}v_t|^2 + (D(v_t), v).\end{aligned}$$

Integrating the last inequality from 0 to T with respect to t we obtain:

$$\begin{aligned}\int_0^T \tilde{E}(s) ds &\leq -(Mv_t(T), v(T)) + (Mv_t(0), v(0)) + \frac{3}{2} \int_0^T |M^{1/2}v_s(s)|^2 ds \\ &\quad - \frac{1}{2} (1 - \epsilon_{\mathcal{L}}^{4\eta} (1 + \delta) b_0) \int_0^T |A^{1/2}v(s)|^2 ds \\ &\quad + \int_0^T (D(v_t(s)), v(s)) ds + C_{\mathcal{L}, \delta} \int_0^T \max_j |l_j(v(s))|^2 ds. \quad (15)\end{aligned}$$

From Lemma 5 we obtain

$$|A^{1/2}v(t)|^2 + |M^{1/2}v(t)|^2 \leq C\tilde{E}(t) + C_{\mathcal{L}} \max_j |l_j(v(t))|^2 \quad (16)$$

under the condition $\epsilon_{\mathcal{L}}^{4\eta} b_0 < 1$. Thus, by direct computations

$$\begin{aligned} |(Mv_t(T), v(T))| + |(Mv_t(0), v(0))| &\leq C[\tilde{E}(T) + \tilde{E}(0)] \\ &+ C_{\mathcal{L}} \left(\max_j |l_j(v(T))|^2 + \max_j |l_j(v(0))|^2 \right). \end{aligned}$$

By (5)

$$\int_0^T (D(v_t), v) ds \leq \int_0^T C(|\mathcal{A}^{1/2}v|) \cdot (D(v_t), v_t) ds + c_2 \int_0^T |\mathcal{A}^{1/2-\eta}v|^2 ds.$$

As above, it is easy to see that $C(|\mathcal{A}^{1/2}v(s)|) \leq C_0 \equiv C_0(E(u_0, u_1), w)$. Therefore, using (13) we obtain

$$\begin{aligned} \int_0^T (D(v_t), v) ds &\leq C_0 \int_0^T (D(v_t), v_t) ds + c_2 \epsilon_{\mathcal{L}}^{4\eta} (1 + \delta) \int_0^T |\mathcal{A}^{1/2}v(s)|^2 ds \\ &+ C_{\mathcal{L}, \delta} \int_0^T \max_j |l_j(v(s))|^2 ds. \end{aligned} \quad (17)$$

From (4) and Jensen's inequality we also have

$$\int_0^T (Mv_t, v_t) ds \leq \int_0^T H((D(v_t), v_t)) ds \leq \frac{T}{3} H_T \left(\int_0^T (D(v_t), v_t) ds \right). \quad (18)$$

Therefore applying inequalities (17) and (18) to main inequality in (15) we obtain

$$\int_0^T \tilde{E}(s) ds \leq C[\tilde{E}(0) + \tilde{E}(T)] + (C_0 I + \frac{T}{2} H_T) \left(\int_0^T (D(v_t), v_t) ds \right) + \mathcal{N}_{\mathcal{L}}(1, T)$$

provided $\epsilon_{\mathcal{L}}^{4\eta}(b_0 + 2c_2) < 1$. Since $t\tilde{E}(t) \leq \int_0^t \tilde{E}(s) ds$ (by (11) $\tilde{E}(t)$ is a decreasing function), we obtain the following inequality

$$T\tilde{E}(T) \leq C_1 \tilde{E}(T) + (C_0 I + \frac{T}{2} H_T) \left(\int_0^T (D(v_t), v_t) ds \right) + \mathcal{N}_{\mathcal{L}}(1, T).$$

Using the relation

$$\int_0^T (D(v_t(s)), v_t(s)) ds = \tilde{E}(0) - \tilde{E}(T),$$

which follows from (11), and taking $T > \max[2C_1, 2]$ we obtain

$$\tilde{E}(T) \leq (C_0 I + H_T) (\tilde{E}(0) - \tilde{E}(T)) + \mathcal{N}_{\mathcal{L}}(1, T).$$

Using the concavity of H_T and recalling the definition of $p(s)$ applied with $c = C_0$ we obtain

$$2p(\tilde{E}(T)/2) + \tilde{E}(T) \leq \tilde{E}(0) + p(\mathcal{N}_{\mathcal{L}}(1, T)) ,$$

which gives the inequality in Lemma 6 for the case $m = 1$. Repeating this argument on each subinterval $((m-1)T, mT)$ gives the desired conclusion in Lemma 6.

In the special case when $H(s) = c_1 s$, $H_T(s) = \tilde{c} s/T$, defining $\gamma < 1$ in an appropriate way we find that $\tilde{E}(T) \leq \gamma \tilde{E}(0) + \mathcal{N}_{\mathcal{L}}(1, T)$, which implies the statement in the second part of Lemma 6 with $m = 1$. Repeating the same argument on the interval $[(m-1)T, mT]$ yields the desired conclusion in the Lemma. ■

To complete the proof of Theorem 3 it suffices to apply Lemma 6 on successive time intervals. This gives, for the case of linear $H(s)$

$$\tilde{E}(nT) \leq \gamma^n \tilde{E}(0) + C_{\mathcal{L}} \sum_{i=1}^n \gamma^i L_{n-i}, \quad (19)$$

where $L_i = \max_{t \in [(m-1)T, mT]} \max_j |l_j v(t)|^2$ and $0 < \gamma < 1$. The conclusion of Theorem 3 for linear $H(s)$ easily follows from (19). In general case we use a comparison theorem similar to Lemma 3.3 [10]. ■

Remark 7 Assume that there exists a positive function $h(t)$ on \mathbf{R}_+ with the properties described in Part II of Theorem 3. In the case of linear $H(s)$ we can derive from (19) the relation

$$\tilde{E}(t) \leq C \tilde{E}(0) e^{-\tilde{\gamma} t} + C_{\mathcal{L}} \int_0^t e^{-\tilde{\gamma}(t-\tau)} h(\tau) d\tau.$$

for some $\tilde{\gamma} > 0$. This formula describes the decay rates of the solution $u(t)$ to the equilibrium w . The corresponding decay rates are exponential when $h(t) = ce^{-\beta t}$ or polynomial if $h(t) = c(1+t)^{-\beta}$, $\beta > 0$. We note that in general $h(t)$ may depend crucially on the solutions $u(t)$ and w .

Theorem 3 implies immediately the following assertion.

Corollary 8 Let $w_1, w_2 \in \mathcal{D}(\mathcal{A}^{1/2})$ be two stationary solutions to (2). Let $\mathcal{L} = \{l_j : j = 1, \dots, N\}$ be a set of functionals on $\mathcal{D}(\mathcal{A}^{1/2})$ and $\epsilon_{\mathcal{L}}^{4\eta}(b_0 + 2c_2) < 1$, where η , b_0 , and c_2 are the same as in Theorem 3. Then the condition $l_j(w_1) = l_j(w_2)$ for $j = 1, \dots, N$ implies that $w_1 = w_2$.

The two corollaries formulated below deal with the case when the problem possesses precompact trajectories (this property holds in the application considered below for the case $\alpha > 0$).

Let $u(t)$ be a solution to the problem (2). We recall that the set

$$\gamma_+(u_0, u_1) = \cup \{(u(t); u_t(t)) : t \geq 0\}$$

in the space $H = \mathcal{D}(\mathcal{A}^{1/2}) \times V$ is said to be the *semi-trajectory* emanating from $(u_0; u_1)$ of the dynamical system generated by (2) in H . We also define the ω -limit set of the semi-trajectory $\gamma_+(u_0, u_1)$ by the formula

$$\omega(\gamma_+) \equiv \omega(u_0, u_1) = \cap_{\tau > 0} [\cup \{(u(t); u_t(t)) : t \geq \tau\}]_H,$$

where $[A]_H$ is the closure of the set A in H .

Corollary 9 *Let $\mathcal{L} = \{l_j : j = 1, \dots, N\}$ be a set of functionals on $\mathcal{D}(\mathcal{A}^{1/2})$ such that $\epsilon_{\mathcal{L}}^{4\eta}(b_0 + 2c_2) < 1$, where η , b_0 , and c_2 are the constants from (5) and (6). Assume that $u(t)$ is a solution to problem (2) with precompact semi-trajectory $\gamma_+ = \gamma_+(u_0, u_1)$ and there exists the finite limits $\lim_{t \rightarrow +\infty} l_j(u(t)) \equiv l_j$ for every $j = 1, \dots, N$. Then there exists a stationary solution $w \in \mathcal{D}(\mathcal{A}^{1/2})$ such that (8) holds.*

Proof. Precompactness of γ_+ implies that ω -limit set $\omega(\gamma_+)$ is non-empty compact set in H . The relation (3) implies that the functional $V(y) = E(u_0, u_1)$, $y = (u_0; u_1)$, is a strict Lyapunov function on H for system (2) (see books [1], [4] or [8] for the definition) and therefore the ω -limit set $\omega(\gamma_+)$ lies in the set $\mathcal{N} \equiv \{(w; 0) \in H : Aw = F(w)\}$ of equilibrium points to problem (2). From the convergence $l_j(u(t)) \rightarrow l_j$ we have that $l_j(w) = l_j$ for all $(w; 0) \in \omega(\gamma_+) \subset \mathcal{N}$. Consequently Corollary 8 implies that $\omega(\gamma_+)$ consists of a single point $(w; 0)$ and therefore (8) holds. ■

Corollary 10 *Let the assumptions of Corollary 9 be valid. Assume that $u^{(1)}(t)$ and $u^{(2)}(t)$ are solutions to equation (2) with precompact semi-trajectories $\gamma_+^{(1)}$ and $\gamma_+^{(2)}$ and*

$$\lim_{t \rightarrow +\infty} (l_j(u^{(1)}(t)) - l_j(u^{(2)}(t))) = 0, \quad j = 1, \dots, N. \quad (20)$$

Then $\omega(\gamma_+^{(1)}) \equiv \omega(\gamma_+^{(2)})$. If the set \mathcal{N} of equilibrium points is finite, then there exists a stationary solution $w \in \mathcal{D}(\mathcal{A}^{1/2})$ such that (8) holds for both solutions $u_1(t)$ and $u_2(t)$.

Proof. Let $(z_1, 0) \in \omega(\gamma_+^{(1)}) \subset \mathcal{N}$. Then there exists a sequence $\{t_m\}$ such that $t_m \rightarrow +\infty$ and $u^{(1)}(t_m) \rightarrow z_1$ in the space $\mathcal{D}(\mathcal{A}^{1/2})$ when $m \rightarrow \infty$. Since $\gamma_+^{(2)}$ is precompact set, we can choose a subsequence $\{t_{m_k}\} \subset \{t_m\}$ such that $u^{(2)}(t_{m_k}) \rightarrow z_2$, where $(z_2, 0) \in \omega(\gamma_+^{(2)}) \subset \mathcal{N}$. Property (20) gives

that $l_j(z_1) = l_j(z_2)$ for all $j = 1, 2, \dots, N$. Consequently Corollary 8 implies that $z_1 = z_2$ and therefore we have $(z_1, 0) \in \omega(\gamma_+^{(2)})$. This implies that $\omega(\gamma_+^{(1)}) = \omega(\gamma_+^{(2)})$. If \mathcal{N} is finite, then it is easy to see that $\omega(\gamma_+^{(1)}) = \omega(\gamma_+^{(2)})$ consists of a single equilibrium point. This implies the assertion of the corollary. ■

Applications

In this section we deal with the following problem

$$\begin{cases} (1 - \alpha \cdot \Delta) \partial_t^2 u + d_0(x) \cdot g_0(u_t) - \alpha \operatorname{div}(d(x)g(\nabla u_t)) \\ + \Delta^2 u - [u, v + F_0] = p(x), & x \in \Omega, t > 0, \\ u|_{\partial\Omega} = \frac{\partial u}{\partial n}|_{\partial\Omega} = 0, \quad u|_{t=0} = u_0(x), \quad \partial_t u|_{t=0} = u_1(x), \end{cases} \quad (21)$$

where $[u, v] = \partial_{x_1}^2 u \cdot \partial_{x_2}^2 v + \partial_{x_2}^2 u \cdot \partial_{x_1}^2 v - 2 \cdot \partial_{x_1 x_2}^2 u \cdot \partial_{x_1 x_2}^2 v$ and $v = v(u)$ is a solution to the problem

$$\Delta^2 v + [u, u] = 0, \quad v|_{\partial\Omega} = \frac{\partial v}{\partial n}|_{\partial\Omega} = 0. \quad (22)$$

Here Ω is a smooth bounded domain in \mathbf{R}^2 , $F_0(x) \in H^4(\Omega)$ and $p(x) \in L_2(\Omega)$ are given functions determined by the mechanical loads. The parameter $\alpha \geq 0$ takes into account the rotational inertial momenta of the elements of the plate. The terms $g_0(u_t)$ and $g(u_t)$ represent mechanical (potentially nonlinear) damping in the system with damping parameters d_0, d nonnegative and bounded in Ω .

We introduce the following spaces and operators:

- $\mathcal{H} \equiv L_2(\Omega)$, $V \equiv H_0^1(\Omega)$ in the case $\alpha > 0$ and $V = \mathcal{H}$ for $\alpha = 0$.
- $\mathcal{A}u \equiv \Delta^2 u$, $u \in \mathcal{D}(\mathcal{A})$ with $\mathcal{D}(\mathcal{A}) \equiv H_0^2(\Omega) \cap H^4(\Omega)$.
- $Mu \equiv Iu - \alpha \Delta u$, $u \in \mathcal{D}(M) \equiv H_0^1(\Omega) \cap H^2(\Omega)$.
- Hence $V' = H^{-1}(\Omega)$ ($\alpha > 0$) and $V' = L_2(\Omega)$ ($\alpha = 0$), $\mathcal{D}(\mathcal{A}^{1/2}) = H_0^2(\Omega)$, $[\mathcal{D}(\mathcal{A}^{1/2})]' = H^{-2}(\Omega)$.
- $F(u) \equiv [u, v(u) + F_0] + p$, where $v(u)$ satisfies (22).
- $D(u) \equiv d_0(x)g_0(u) - \alpha \operatorname{div}(d(x)g(\nabla u))$.

The nonlinear term $F(u)$ has the form $F(u) = -\Pi'(u)$ with

$$\Pi(u) = \frac{1}{4} \|\Delta v(u)\|^2 - \frac{1}{2} ([u, u], F_0) - ([p, u],$$

where $v(u) \in H_0^2(\Omega)$ is defined by (22).

We assume that the mappings $g_0 : \mathbf{R} \mapsto \mathbf{R}$ and $g : \mathbf{R}^2 \mapsto \mathbf{R}^2$ are locally Lipschitz and possess the properties $g(s_1, s_2) = (g_1(s_1); g_2(s_2))$ for $(s_1; s_2) \in \mathbf{R}^2$, and $g_i(0) = 0$, $i = 0, 1, 2$, and $g_i(s)$ is non-decreasing for each $i = 0, 1, 2$.

Under these conditions one can show (see, e.g., [6], [12] or [13]) that Assumption 1 holds for both $\alpha > 0$ and $\alpha = 0$ cases. Thus problem (21) and (22) has a unique solution $u(t)$ in $C(\mathbf{R}_+; H_0^2(\Omega)) \cap C(\mathbf{R}_+; V_\alpha(\Omega))$, where $V_\alpha(\Omega) = H_0^1(\Omega)$ for $\alpha > 0$ and $V_\alpha(\Omega) = L_2(\Omega)$ for $\alpha = 0$.

To check assumption (A5) we need some additional hypotheses concerning $d_0(x)g_0(v)$ and $d(x)g(v)$. We assume that (a) there exist positive constants d_0 and d_1 such that $d_0 \leq d_0(x) \leq d_1$ and $d_0 \leq d(x) \leq d_1$ and (b) there exist positive constants a, b and $q \geq 1$ such that

$$sg_i(s) \geq as^2, \quad i = 0, 1, 2, |s| \geq 1; \quad |g_i(s)| \leq b(1 + |s|^q), \quad i = 1, 2. \quad (23)$$

As for assumption (A6), we can show that

$$\begin{aligned} (F(w + u) - F(w + z \cdot u), u) &\leq -\frac{1-z}{2} \|\Delta v(u)\|^2 \\ &+ c_0(1-z) \|u\|_{H^1(\Omega)}^2 (\|\Delta w\|^2 + \|F_0\|_{H^4(\Omega)}^2) \end{aligned}$$

for any $z \in [0, 1]$ and $u, w \in H_0^2(\Omega)$, where c_0 is a constant depending on embedding theorems. Since the set of stationary solutions to (21) is bounded (see, e.g., [6]), the assumption (A6) holds in this case.

Remark 11 In the case considered above, when we assume that the dissipation g_i , $i = 0, 1, 2$, is differentiable near the origin, we can make the corresponding constant c_2 arbitrary small. We also have $\eta = 3/8$ and $b_0 = c_0 (\|\Delta w\|^2 + \|F_0\|_{H^4(\Omega)}^2)$ for the system under consideration.

We also note that if (23) holds for all $s \in \mathbf{R}$, then $H(s) = c_1 s$ with a suitable c_1 . In general, construction of appropriate function $H(s)$ relies on monotonicity of $g(s)$ and asymptotic growth condition in (23) and can be done in the same way as in [10] for wave equation.

Under all these assumptions Theorem 3 is applicable here and we have the following assertion.

Theorem 12 Let $u(t)$ be a solution to problem (21) and (22) and $w \in H_0^2(\Omega)$ be a solution to the problem

$$\Delta^2 u - [u, v + F_0] = p(x), \quad x \in \Omega, \quad u|_{\partial\Omega} = \frac{\partial u}{\partial n}|_{\partial\Omega} = 0,$$

where v solves (22). Assume that $\mathcal{L} = \{l_j : j = 1, \dots, N\}$ is a set of functionals on $H_0^2(\Omega)$ with the completeness defect $\epsilon_{\mathcal{L}} \equiv \epsilon_{\mathcal{L}}(H_0^2(\Omega), L_2(\Omega))$.

Then there exists $\epsilon_0 > 0$ such that the conditions $\epsilon_{\mathcal{L}} < \epsilon_0$ and

$$\lim_{t \rightarrow +\infty} l_j(u(t)) = l_j(w) \quad \text{for all } j = 1, \dots, N$$

imply that

$$\lim_{t \rightarrow +\infty} \left(\|u_t(t)\|_{L^2(\Omega)}^2 + \alpha \|\nabla u_t(t)\|_{L^2(\Omega)}^2 + \|\Delta(u(t) - w)\|_{L^2(\Omega)}^2 \right) = 0.$$

Remark 13 Theorem 12 implies the property of stationary solutions to von Karman equations which is similar to Corollary 8. However the precompactness of trajectories of the dynamical system generated by (21) and (22) we can guarantee in the case $\alpha > 0$ only. Therefore this is only the case when Corollaries 9 and 10 can be applied. For details we refer to [6]. We also note that for the case $\alpha = 0$, $g_0(s) = g_0 \cdot s$ determining functionals for problem (21) and (22) were studied in [2].

The situation when the completeness defect $\epsilon_{\mathcal{L}}$ can be easily estimated from above is the following. Assume that $\{\phi_j : j = 1, \dots, N\}$ be a certain set of linearly independent functions from $H_0^2(\Omega)$. Let us define an interpolation operator $R_{\mathcal{L}}$ by the formula $R_{\mathcal{L}}w = \sum_{j=1}^N l_j(w)\phi_j$. If $R_{\mathcal{L}}w$ is a "good" approximation for w , we have

$$\|w - R_{\mathcal{L}}w\|_{L_2(\Omega)} \leq Ch^\alpha \|w\|_{H^2(\Omega)},$$

where C and α are positive constants and $h > 0$ is small enough. In this case we obviously have $\epsilon_{\mathcal{L}} \leq Ch^\alpha$. This observation allows us to give the following examples (for details and further discussion we refer to [3, 4]).

Local volume averages. Assume that $\lambda(x) \in L^\infty(\mathbf{R}^2)$ has compact support and $\int_{\mathbf{R}^2} \lambda(x) dx = 1$. For $h > 0$ let us define functionals

$$\mathcal{L} = \left\{ l_j : l_j(w) = \frac{1}{h^2} \cdot \int_{\Omega} w(x) \lambda\left(\frac{x}{h} - j\right) dx, \quad j \equiv (j_1, j_2) \in \mathcal{J} \right\},$$

where $\mathcal{J} \equiv \{(j_1, j_2) \in \mathbf{Z}^2 : (j_1 h, j_2 h) \in \Omega\}$. In this case we have the estimate $\epsilon_{\mathcal{L}}(H_0^2(\Omega), L_2(\Omega)) \leq ch^2$.

Nodes. Let \mathcal{T}^h be a triangulation of the domain Ω , made of triangles with sides less than h and let $\{x_j : j = 1, \dots, N_h\}$ be the set of all vertices of the triangles from \mathcal{T}^h . Then the completeness defect for the set

$$\mathcal{L} = \{l_j : l_j(w) = w(x_j), \quad j = 1, \dots, N_h\}$$

admits the estimate $\epsilon_{\mathcal{L}}(H_0^2(\Omega), L_2(\Omega)) \leq ch^2$.

Modes. Let $\{e_k\}$ be the basis in $L^2(\Omega)$ consisting of the eigenvectors of the biharmonic operator Δ^2 with the Dirichlet boundary conditions. We suppose

$$\mathcal{L} = \left\{ l_j : l_j(w) = \int_{\Omega} w(x) \cdot e_j(x) dx, \quad j = 1, \dots, N \right\}.$$

Then the estimate $\epsilon_{\mathcal{L}}(H_0^2(\Omega), L_2(\Omega)) \leq cN^{-1}$ holds.

References

- [1] A.V.Babin and M.I.Vishik. *Attractors of Evolution Equations*. North-Holland, Amsterdam, 1992.
- [2] I.Chueshov. On the finiteness of the number of determining elements for von Karman evolution equations. *Math. Meth. in the Appl. Sci.*, 20:855–865, 1997.
- [3] I.Chueshov. Theory of functionals that uniquely determine asymptotic dynamics of infinite-dimensional dissipative systems. *Russian Math. Surveys*, 53:731–776, 1998.
- [4] I.Chueshov. *Introduction to the Theory of Infinite-Dimensional Dissipative Systems*. Acta, Kharkov, 1999 (in Russian); English translation: Acta, Kharkov, 2002; see also <http://www.emis.de/monographs/Chueshov/>
- [5] I.Chueshov and V.Kalantarov. Determining functionals for nonlinear damped wave equations *Matem. Fizika, Analyz, Geometriya*, 8:215–227, 2001.
- [6] I.Chueshov and I.Lasiecka. *Von Karman Evolution Equations*, book in preparation.
- [7] C.Foias and G.Prodi. Sur le comportement global des solutions nonstationnaires des équations de Navier-Stokes en dimension deux. *Rend. Sem. Mat. Univ. Padova*, 39:1–34, 1967.
- [8] J.K.Hale. *Asymptotic Behavior of Dissipative Systems*. AMS, Providence, RI, 1988.
- [9] W.Heyman and I.Lasiecka. Asymptotic behaviour of solutions to nonlinear shells in a supersonic flow. *Numer. Funct. Anal. and Optimization*, 20:279–300, 1999.
- [10] I.Lasiecka and D.Tataru. Uniform boundary stabilization of semilinear wave equation with nonlinear boundary damping. *Int. Diff. Equations*, 6:507–533, 1993.
- [11] O.A.Ladyzhenskaya. A dynamical system generated by the Navier–Stokes equations. *J. Soviet Math.*, 3:458–479, 1975.
- [12] I. Lasiecka. Finite dimensionality and compactness of attractors for von Karman equations with nonlinear dissipation. *Nonlin. Diff. Equations*, 6:437–472, 1999.
- [13] I.Lasiecka. *Mathematical Control Theory of Coupled PDE's*. CMBS-NSF Lecture Notes. SIAM Publications, 2001.

ON SOME CASES OF CAUCHY PROBLEM

Silvia–Otilia Corduneanu

Department of Mathematics

Gh. Asachi Technical University of Iași

11 Copou Blvd., RO-6600 Iași, Romania

silvcord@math.tuiasi.ro

Abstract The study of almost periodic measures depending on the parameter $t \in \mathbb{R}$ suggested to consider some cases of Cauchy problem. The solutions are functions belonging to $\mathcal{C}^1(\mathbb{R}, ap(G))$, where $ap(G)$ is the space of almost periodic measures on a locally compact abelian group G . A generalization of these cases is given by replacing $ap(G)$ with a locally convex space.

Keywords: Cauchy problem, almost periodic measure, almost periodic function.

1. Introduction

Let X be a locally convex space, Y a Banach algebra and $\bullet : Y \times X \rightarrow X$ a bilinear mapping. Consider $y \in Y$ and the linear operator $A : X \rightarrow X$, where $Ax = y \bullet x$, $x \in X$. In our paper we give hypotheses for solving the Cauchy problem

$$\begin{cases} \frac{du(t)}{dt} = Au(t), & t \in \mathbb{R}, \\ u(0) = u_0, \end{cases} \quad (1.1)$$

where $u_0 \in X$ and $u \in \mathcal{C}^1(\mathbb{R}, X)$. We also give some applications for functions which have values in the space of almost periodic measures $ap(G)$. The set $ap(G)$, of all almost periodic measures on a locally compact abelian group G , is a locally convex space with respect to a topology which is called *the product topology*. If ν is a bounded measure and μ is an almost periodic measure, their convolution, $\nu * \mu$, is also an almost periodic measure. These considerations allow us to discuss the Cauchy problem

$$\begin{cases} \frac{du(t)}{dt} = \nu * u(t), & t \in \mathbb{R}, \\ u(0) = u_0. \end{cases} \quad (1.2)$$

Here $u \in \mathcal{C}^1(\mathbb{R}, ap(G))$, ν is a bounded measure on G and $u_0 \in ap(G)$.

2. Preliminaries

Consider a Hausdorff locally compact abelian group G and let λ be the Haar measure on G . Let us denote by $\mathcal{C}(G)$ the set of all bounded continuous complex-valued functions on G , and by $\mathcal{C}_U(G)$ the subset of $\mathcal{C}(G)$ containing the uniformly continuous functions. The sets $\mathcal{C}(G)$ and $\mathcal{C}_U(G)$ are Banach algebras endowed with the supremum norm. Throughout this paper, $\|\cdot\|_u$ denotes the supremum norm on $\mathcal{C}(G)$. For $f \in \mathcal{C}(G)$ and $a \in G$, the translate of f by a is the function $f_a(x) = f(xa)$ for all $x \in G$. Denote by $K(G)$ the linear space of all continuous complex-valued functions on G , having a compact support. We denote by $m(G)$ the space of complex Radon measures on G . That is, the space of all complex linear functionals μ on $K(G)$ satisfying the following: for each compact subset A of G there exists a positive number $m_{\mu,A}$ such that $|\mu(f)| \leq m_{\mu,A} \|f\|_u$ whenever $f \in K(G)$ and the support of f is contained in A . We use $m_F(G)$ to denote the subspace of $m(G)$ consisting of all bounded measures, i.e. all linear functionals which are continuous with respect to the supremum norm on $K(G)$. The action of a measure $\mu \in m(G)$ on a function $f \in K(G)$ will be denoted either $\mu(f)$ or $\int_G f(x)d\mu(x)$. Corresponding to a measure $\mu \in m(G)$, one defines the variation measure $|\mu| \in m(G)$ by $|\mu|(f) = \sup\{|\mu(g)| : g \in K(G), |g| \leq f\}$ for all $f \in K(G), f \geq 0$. For the Borel functions f, g and the measures $\mu, \nu \in m(G)$, we can define their convolutions $f * g$, $f * \mu$, $\nu * \mu$, when that is possible. A translation-bounded measure is a measure $\mu \in m(G)$ with the property that for every compact set $A \subseteq G$, $m_\mu(A) = \sup_{x \in G} |\mu|(xA) < \infty$ (L. N. Argabright and J. G. Lamadrid [1974]). The linear space of the translation-bounded measures will be denoted by $m_B(G)$. We identify an arbitrary measure $\mu \in m_B(G)$ with an element of the space $[\mathcal{C}_U(G)]^{K(G)}$ in the following way: $\mu \equiv \{f * \mu\}_{f \in K(G)}$. From this identification we have the inclusion $m_B(G) \subset [\mathcal{C}_U(G)]^{K(G)}$. The space $[\mathcal{C}_U(G)]^{K(G)}$ has the *product topology* defined by the Banach space structure on $\mathcal{C}_U(G)$, hence, $m_B(G)$ is a locally convex space of measures with the relative topology. A system of seminorms for the product topology on $m_B(G)$ is given by the family $\{\|\cdot\|_f\}_{f \in K(G)}$, where, for a function $f \in K(G)$, $\|\mu\|_f = \|f * \mu\|_u$, for all $\mu \in m_B(G)$. Next we give the definition of an almost periodic function (see E. Hewitt and K. A. Ross [1963]).

Definition 2.1 A function $g \in \mathcal{C}(G)$ is called an almost periodic function on G , if the family of translates of g , $\{g_a : a \in G\}$ is relatively compact in the sense of uniform convergence on G .

The set $AP(G)$ of all almost periodic functions on G is a Banach algebra with respect to the supremum norm, closed to conjugation. Denote

by \hat{G} the dual of G and by $[\hat{G}]$ the linear space generated by \hat{G} in $\mathcal{C}(G)$. It is easy to see that $[\hat{G}] \subset AP(G)$. The almost periodic measures are introduced and studied by L. N. Argabright and J. G. Lamadrid (J. G. Lamadrid [1973], L. N. Argabright and J. G. Lamadrid [1990]).

Definition 2.2 The measure $\mu \in m_B(G)$ is said to be an almost periodic measure, if for every $f \in K(G)$, $f * \mu \in AP(G)$.

The set $ap(G)$ of all almost periodic measures is a locally convex space with respect to the product topology. If $\nu \in m_F(G)$, $\mu \in ap(G)$ then $\nu * \mu \in ap(G)$. Also, if $f \in AP(G)$ and $\mu \in ap(G)$, then the measure $f\mu$, defined by $f\mu(g) = \mu(gf)$, $g \in K(G)$ is an almost periodic measure (see J. G. Lamadrid [1973]). It is also proved that there exists a unique linear functional $M : ap(G) \rightarrow C$ such that M is continuous on $ap(G)$, $M(\mu) = M(\delta_x * \mu)$, for all $x \in G$, $\mu \in ap(G)$ and $M(\lambda) = 1$ (see J. G. Lamadrid [1973]). We used the notation δ_x for the the Dirac measure at $x \in G$. If $\mu \in ap(G)$ we define the mean value of μ as being the above complex number $M(\mu)$.

3. A Cauchy problem

The framework of this section has been suggested by some concrete cases of the Cauchy problem, which are presented in the next section. Let X be a locally convex space with a sufficient family of seminorms \mathcal{P} , $(Y, \|\cdot\|)$ a commutative Banach algebra with the unit element denoted by θ . We also consider the Banach algebra $\mathcal{B}(Y)$ of bounded linear operators on Y and a bilinear mapping $\bullet : Y \times X \rightarrow X$ such that

$$\left\{ \begin{array}{l} (C_1) \quad \theta \bullet x = x, \quad \forall x \in X, \\ (C_2) \quad y \bullet (z \bullet x) = yz \bullet x, \quad \forall y, z \in Y, \quad \forall x \in X, \\ (C_3) \quad \forall p \in \mathcal{P}, \quad p(y \bullet x) \leq \|y\| p(x), \quad \forall y \in Y, \quad \forall x \in X. \end{array} \right.$$

The following proposition contains simple consequences of the properties regarding the series in commutative Banach algebras.

Proposition 3.1 Consider $y \in Y$. Then for every $t \in \mathbb{R}$, the sequence $\theta + t \frac{y}{1!} + \cdots + t^n \frac{y^n}{n!}$ is a Cauchy sequence in the Banach space Y and the limit, which is denoted by e^{ty} , has the following properties:

- (a) $e^{ty} = \theta$ for $t = 0$;
- (b) $e^{(t+s)y} = e^{ty}e^{sy}$ for $t, s \in \mathbb{R}$;
- (c) $\lim_{t \rightarrow 0} \|e^{ty} - \theta\| = 0$;
- (d) $\lim_{t \rightarrow 0} \left\| \frac{e^{ty} - \theta}{t} - y \right\| = 0$.

Next we consider $y \in Y$ and we associate to y the bounded linear operator $A \in \mathcal{B}(Y)$, $Az = yz$, $z \in Y$ and the family of bounded linear operators depending on parameter $t \in \mathbb{R}$, $T(t)$, where for all $t \in \mathbb{R}$, $T(t)z = e^{ty}z$, $z \in Y$. Then $T(t)$, $t \in \mathbb{R}$ is a group of bounded linear operators on Y and A is its generator.

Proposition 3.2 *Consider the Cauchy problem*

$$\begin{cases} \frac{du(t)}{dt} = yu(t), & t \in \mathbb{R}, \\ u(0) = u_0, \end{cases} \quad (3.3)$$

where $y, u_0 \in Y$ and $u \in \mathcal{C}^1(\mathbb{R}, Y)$ is the unknown function. Then there exists a unique solution and that is $u(t) = e^{ty}u_0$, $t \in \mathbb{R}$.

Proof. We observe that if we consider the function $A : Y \rightarrow Y$, $A(z) = yz$ we have that $A \in \mathcal{B}(Y)$, therefore, according to the general theory (see J. A. Goldstein [1985]), the solution of (3.3) is $u(t) = T(t)u_0$, $t \in \mathbb{R}$, where $T(t)$, $t \in \mathbb{R}$ is the group associated to A . In fact we can represent the solution in the form $u(t) = e^{ty}u_0$, $t \in \mathbb{R}$. ■

A function $u \in \mathcal{C}^1(\mathbb{R}, X)$ is called solution of the Cauchy problem

$$\begin{cases} \frac{du(t)}{dt} = y \bullet u(t), & t \in \mathbb{R}, \\ u(0) = u_0, \end{cases} \quad (3.4)$$

if u satisfies the differential equation and the initial condition, which form this problem. The derivative $\frac{du}{dt}(t)$ of the function $u : \mathbb{R} \rightarrow X$ is defined taking the limit of differences quotient in the topology of the locally convex space X . We say that $u \in \mathcal{C}^1(\mathbb{R}, X)$ if there exists the derivative $\frac{du}{dt} : \mathbb{R} \rightarrow X$ and this is a continuous function.

Lemma 3.1 *Consider the functions $u \in \mathcal{C}^1(\mathbb{R}, Y)$, $v \in \mathcal{C}^1(\mathbb{R}, X)$. Then we have the formula*

$$\frac{d}{dt}[u(t) \bullet v(t)] = u'(t) \bullet v(t) + u(t) \bullet v'(t), \quad t \in \mathbb{R}. \quad (3.5)$$

Proof. Let $p \in \mathcal{P}$ and $t \in \mathbb{R}$. Using (C_3) it follows that for all $h \in \mathbb{R}$, $h \neq 0$ we have

$$\begin{aligned} & p \left[\frac{u(t+h) \bullet v(t+h) - u(t) \bullet v(t)}{h} - u'(t) \bullet v(t) - u(t) \bullet v'(t) \right] \\ & \leq \left\| \frac{u(t+h) - u(t)}{h} - u'(t) \right\| p[v(t+h)] \\ & + \|u(t)\| p \left[\frac{v(t+h) - v(t)}{h} - v'(t) \right] + \|u'(t)\| p[v(t+h) - v(t)]. \end{aligned}$$

Therefore we obtain

$$\lim_{h \rightarrow 0} p \left[\frac{u(t+h) \bullet v(t+h) - u(t) \bullet v(t)}{h} - u'(t) \bullet v(t) - u(t) \bullet v'(t) \right] = 0.$$

Consequently, formula (3.5) has been proved. ■

Theorem 3.1 Let $y \in Y$ and consider the family of linear operators on X , $T(t)$, $t \in \mathbb{R}$, where for every $t \in \mathbb{R}$, $T(t)x = e^{ty} \bullet x$, $x \in X$. Then for all $x \in X$ we have that

$$\frac{d}{dt} T(t)x = y \bullet T(t)x, \quad t \in \mathbb{R}. \quad (3.6)$$

Proof. Let p be in \mathcal{P} , $t \in \mathbb{R}$ and $x \in X$. For all $h \in \mathbb{R}$, $h \neq 0$ we have

$$\begin{aligned} p \left(\frac{T(t+h)x - T(t)x}{h} - y \bullet T(t)x \right) &= p \left(\frac{e^{(t+h)y} \bullet x - e^{ty} \bullet x}{h} - ye^{ty} \bullet x \right) \\ &\leq \left\| \frac{e^{(t+h)y} - e^{ty}}{h} - ye^{ty} \right\| p(x) \leq \|e^{ty}\| \left\| \frac{e^{hy} - \theta}{h} - y \right\| p(x). \end{aligned}$$

Therefore

$$\lim_{h \rightarrow 0} p \left(\frac{T(t+h)x - T(t)x}{h} - y \bullet T(t)x \right) = 0, \text{ for all } p \in \mathcal{P}$$

and this shows (3.6). ■

Corollary 3.1 Consider $y \in Y$, $u_0 \in X$ and the Cauchy problem

$$\begin{cases} \frac{du(t)}{dt} = y \bullet u(t), & t \in \mathbb{R}, \\ u(0) = u_0, \end{cases} \quad (3.7)$$

where $u \in \mathcal{C}^1(\mathbb{R}, X)$ is the unknown function. Then (3.7) has a unique solution $u \in \mathcal{C}^1(\mathbb{R}, X)$ and $u(t) = e^{ty} \bullet u_0$, $t \in \mathbb{R}$.

Proof. From Theorem 3.1 it follows that the function $u(t) = e^{yt} \bullet u_0$, $t \in \mathbb{R}$ verifies the differential equation contained in (3.7) and from Proposition 3.1 it results that this function satisfies the initial condition. Consider $t > 0$. We now prove that if $u \in \mathcal{C}^1(\mathbb{R}, X)$ is a solution of (3.7)

then the function $s \in [0, t] \rightarrow e^{(t-s)y} \bullet u(s) \in X$ has the property that

$$\frac{d}{ds}[e^{(t-s)y} \bullet u(s)] = 0_X, \quad s \in [0, t]. \quad (3.8)$$

From Lemma 3.1 and Proposition 3.1 it results that

$$\frac{d}{ds}[e^{(t-s)y} \bullet u(s)] = -ye^{(t-s)y} \bullet u(s) + e^{(t-s)y} \bullet u'(s), \quad s \in [0, t]. \quad (3.9)$$

Combining (3.7) and (3.9) we obtain (3.8). Taking into account that \mathcal{P} is a sufficient family of seminorms on X it results that function $s \in [0, t] \rightarrow e^{(t-s)y} \bullet u(s) \in X$ is constant. The equality $u(0) = u(t)$ gives us that $u(t) = e^{ty} \bullet u_0$, and from the fact that $t > 0$ is arbitrary we obtain $u(t) = e^{ty} \bullet u_0$, $t \in [0, \infty)$. Finally, similar arguments for $t < 0$ lead us to the conclusion that $u(t) = e^{ty} \bullet u_0$, $t \in \mathbb{R}$. ■

4. Particular cases of Cauchy problem

In this section we present some particular cases of the Cauchy problem (3.7). In the first case Y is the Banach algebra $m_F(G)$ of bounded measures on a locally compact abelian group G . We remind that $m_F(G)$ is a commutative Banach algebra with respect the usual norm of bounded measures which is denoted by $\|\cdot\|$. The third operation of $m_F(G)$ is the convolution of bounded measures and the unit element is the Dirac measure $\delta_e \in m_F(G)$, e being the unit element of the group G . On the other hand X is the locally convex space $(ap(G), \{\|\cdot\|_f\}_{f \in K(G)})$ of all almost periodic measures on G . The bilinear mapping $\bullet : Y \times X \rightarrow X$ is given by the convolution between a bounded measure and an almost periodic measure which is an almost periodic measure. We start by giving the following lemma which will play an important role in proving Corollary 4.1.

Lemma 4.1 *If $\nu \in m_F(G)$, $\mu \in ap(G)$ and $f \in K(G)$ we have that*

$$\|\nu * \mu\|_f \leq \|\mu\|_f \|\nu\|.$$

Proof. We notice that

$$\|\nu * \mu\|_f = \|f * \nu * \mu\|_u = \sup_{x \in G} \left| \int_G f * \mu(xy^{-1}) d\nu(y) \right|.$$

Consider $x \in G$. Well-known properties of the integral yield

$$\left| \int_G f * \mu(xy^{-1}) d\nu(y) \right| \leq \int_G |f * \mu|(xy^{-1}) d|\nu|(y) \leq \|f * \mu\|_u \|\nu\| = \|\mu\|_f \|\nu\|.$$

Hence $\|\nu * \mu\|_f \leq \|\mu\|_f \|\nu\|$. ■

Corollary 4.1 Consider $\nu \in m_F(G)$, $u_0 \in ap(G)$ and the Cauchy problem

$$\begin{cases} \frac{du(t)}{dt} = \nu * u(t), & t \in \mathbb{R}, \\ u(0) = u_0, \end{cases} \quad (4.10)$$

where $u \in C^1(\mathbb{R}, ap(G))$ is an unknown function. Then (4.10) has a unique solution and this is $u \in C^1(\mathbb{R}, ap(G))$, $u(t) = e^{t\nu} * u_0$, $t \in \mathbb{R}$.

Proof. The properties of the convolution between a bounded measure and an almost periodic measure and Lemma 4.1 enable us to see that the bilinear mapping $* : m_F(G) \times ap(G) \rightarrow ap(G)$, satisfies conditions (C_1) , (C_2) , (C_3) from the preceding section. So, the Cauchy problem (4.10) is a particular case of (3.7). In what follows we can apply Corollary 3.1 and we find the conclusion. ■

Remark 4.1 In the previous corollary we can replace the locally convex space $ap(G)$ by the Banach space $AP(G)$ and we obtain, in the same manner, a similar result. So, if $\nu \in m_F(G)$, $u_0 \in AP(G)$, then the Cauchy problem

$$\begin{cases} \frac{du(t)}{dt} = \nu * u(t), & t \in \mathbb{R}, \\ u(0) = u_0, \end{cases} \quad (4.11)$$

has a unique solution and this is $u \in C^1(\mathbb{R}, AP(G))$, $u(t) = e^{t\nu} * u_0$, $t \in \mathbb{R}$.

In the second case Y is the commutative Banach algebra $AP(G)$ of all almost periodic functions on G and X is the locally convex space of all almost periodic measures on G , $(ap(G), \{\|\cdot\|\}_{f \in K(G)})$. The unit element of the $AP(G)$ is the constant function denoted by $\mathbf{1}$, which takes the value 1 for all $x \in G$. This time, the bilinear mapping associates to an almost periodic function $g \in AP(G)$ and an almost periodic measure $\mu \in ap(G)$ the almost periodic measure having the density g and the base μ .

Lemma 4.2 If $g \in AP(G)$, $\mu \in ap(G)$ and $f \in K(G)$ we have that

$$\|g\mu\|_f \leq \|g\|_u \| |f| * |\mu| \|_u.$$

Proof. We have that $\|g\mu\|_f = \|f * g\mu\|_u = \sup_{x \in G} \left| \int_G f(xy^{-1}) g(y) d\mu(y) \right|$.

Consider $x \in G$. Clearly, the following inequalities hold

$$\begin{aligned} \left| \int_G f(xy^{-1}) g(y) d\mu(y) \right| &\leq \int_G |f(xy^{-1})| |g(y)| d|\mu|(y) \\ &\leq \|g\|_u \int_G |f(xy^{-1})| d|\mu|(y) \leq \|g\|_u \| |f| * |\mu| \|_u. \end{aligned}$$

Therefore $\|g\mu\|_f \leq \|g\|_u \|\cdot f \cdot |\mu|\|_u$. ■

Corollary 4.2 Consider $g \in AP(G)$, $u_0 \in ap(G)$ and the Cauchy problem

$$\begin{cases} \frac{du(t)}{dt} = gu(t), & t \in \mathbb{R}, \\ u(0) = u_0, \end{cases} \quad (4.12)$$

where $u \in C^1(\mathbb{R}, ap(G))$ is an unknown function. Then (4.12) has a unique solution $u \in C^1(\mathbb{R}, ap(G))$ and $u(t) = e^{tg}u_0$, $t \in \mathbb{R}$.

Proof. We consider the bilinear mapping which associates to an almost periodic function $g \in AP(G)$ and an almost periodic measure $\mu \in ap(G)$ the almost periodic measure having the density g and the base μ . It is easy to see that this bilinear mapping verifies (C_1) , (C_2) but does not satisfy (C_3) . However, there exists a unique solution and this is $u \in C^1(\mathbb{R}, ap(G))$, $u(t) = e^{tg}u_0$, $t \in \mathbb{R}$. First let us see that, by Proposition 3.1 and Lemma 4.2 it results that for all $f \in K(G)$, $t \in \mathbb{R}$, $h \neq 0$ we have

$$\left\| \frac{e^{(t+h)g}u_0 - e^{tg}u_0}{h} - ge^{tg}u_0 \right\|_f \leq \|e^{tg}\|_u \left\| \frac{e^{hg} - 1}{h} - g \right\|_u \|\cdot f \cdot |u_0|\|_u.$$

Hence, for all $t \in \mathbb{R}$ and $f \in K(G)$ we obtain that

$$\lim_{h \rightarrow 0} \left\| \frac{e^{(t+h)g}u_0 - e^{tg}u_0}{h} - ge^{tg}u_0 \right\|_f = 0$$

and this means that $u(t) = e^{tg}u_0$, $t \in \mathbb{R}$ verifies the differential equation contained in (4.12). By virtue of the properties of the variation measure and taking into consideration Lemma 4.2 we can establish that for a fixed $t > 0$ we have

$$\frac{d}{ds} [e^{(t-s)g}u(s)] = 0_{ap(G)}, \quad s \in [0, t].$$

Here, we have denoted by $0_{ap(G)}$, the null measure on G . Thus, similar arguments as in the proof of Corollary 3.1 lead us to the conclusion $u(t) = e^{tg}u_0$, $t \in \mathbb{R}$ is the unique solution of (4.12). ■

5. Examples and applications

To illustrate the technique of finding the solution for the Cauchy problems (4.10) and (4.11), we consider some examples. The solution of our first example is a function $u \in C^1(\mathbb{R}, AP(\mathbb{R}))$.

Example 5.1 Consider $f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, $x \in \mathbb{R}$, $u_0(x) = \cos x$, $x \in \mathbb{R}$ and $\nu = f\lambda$ where λ is the Lebesgue measure on \mathbb{R} . It is obvious that $\nu \in m_F(\mathbb{R})$ and $u_0 \in AP(\mathbb{R})$. With these notations (4.11) becomes

$$\begin{cases} \frac{du(t)}{dt}(x) = \int_{\mathbb{R}} u(t)(x-y) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy, & x, t \in \mathbb{R}, \\ u(0)(x) = \cos x, & x \in \mathbb{R}, \end{cases} \quad (5.13)$$

where $u \in \mathcal{C}^1(\mathbb{R}, AP(\mathbb{R}))$. For all $t \in \mathbb{R}$ we have that

$$e^{t\nu} = \delta_0 + t \frac{f\lambda}{1!} + \cdots + t^n \frac{f^{[n]}\lambda}{n!} + \cdots,$$

where $f^{[n]} = f * f * \cdots * f$. Some calculations give us that

$$f^{[n]}(x) = \frac{1}{\sqrt{2n\pi}} e^{-\frac{x^2}{2n}}, \quad x \in \mathbb{R}, \quad n \in \mathbb{N}^*,$$

$$f^{[n]}\lambda * u_0(x) = \frac{1}{\sqrt{2n\pi}} \int_{\mathbb{R}} \cos(x-y) e^{-\frac{y^2}{2n}} dy = e^{-\frac{n}{2}} \cos x, \quad x \in \mathbb{R}, \quad n \in \mathbb{N}^*.$$

Finally we obtain

$$u(t)(x) = \cos x + \sum_{n=1}^{\infty} \frac{e^{-\frac{n}{2}t^n}}{n!} \cos x = e^{\frac{t}{\sqrt{e}}} \cos x, \quad x, t \in \mathbb{R}.$$

In the second Cauchy problem, the solution is a function $u \in \mathcal{C}^1(\mathbb{R}, ap(\mathbb{R}))$.

Example 5.2 Consider $f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, $x \in \mathbb{R}$, $u_0 = \cos(\cdot)\lambda$ and $\nu = f\lambda$, where λ is the Lebesgue measure on \mathbb{R} . It is obvious that $\nu \in m_F(\mathbb{R})$ and $u_0 \in ap(\mathbb{R})$. With these notations (4.10) becomes

$$\begin{cases} \frac{du(t)}{dt}(\varphi) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \varphi(x+y) d[u(t)](x) \right) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy, & t \in \mathbb{R}, \\ u(0) = \cos(\cdot)\lambda, \end{cases}$$

where $\varphi \in K(\mathbb{R})$ and $u \in \mathcal{C}^1(\mathbb{R}, ap(\mathbb{R}))$. In a similar manner by that used in Example 5.1 we obtain that

$$u(t) = \left[e^{\frac{t}{\sqrt{e}}} \cos(\cdot) \right] \lambda, \quad t \in \mathbb{R}.$$

Finally, we establish equalities for the mean of some classes of almost periodic measures depending on the parameter $t \in \mathbb{R}$. Let $u \in \mathcal{C}^1(\mathbb{R}, ap(G))$. It is easy to see that the function $t \in \mathbb{R} \rightarrow M[u(t)] \in \mathbb{C}$ satisfies the equality

$$\frac{d}{dt} M[u(t)] = M \left[\frac{du}{dt}(t) \right], \quad t \in \mathbb{R}. \quad (5.14)$$

Next, we apply (5.14) to the solutions of the Cauchy problems (4.10) and (4.12). Consider $\nu \in m_F(G)$, $\mu \in ap(G)$ such that $M(\mu) \neq 0$ and $g \in AP(G)$. First, from (5.14) we deduce the equality

$$\frac{d}{dt} M[e^{t\nu} * \mu] = [\nu(G)] M[e^{t\nu} * \mu], \quad t \in \mathbb{R}. \quad (5.15)$$

We use a property of the mean, respectively, the equality

$$M(e^{t\nu} * \mu) = e^{t\nu}(G)M(\mu), \quad t \in \mathbb{R},$$

and we obtain that the following equality holds true:

$$\frac{d}{dt}[e^{t\nu}(G)] = \nu(G)e^{t\nu}(G), \quad t \in \mathbb{R}. \quad (5.16)$$

On the other hand, from (5.14), it also follows that

$$\frac{d}{dt}M[e^{tg}\mu] = M[ge^{tg}\mu], \quad t \in \mathbb{R}.$$

In the same manner we obtain that for all $n \in \mathbb{N}^*$ we have,

$$\begin{aligned} \frac{d^n}{dt^n}M[e^{t\nu} * \mu] &= [\nu(G)]^n M[e^{t\nu} * \mu], \quad t \in \mathbb{R}, \\ \frac{d^n}{dt^n}[e^{t\nu}(G)] &= [\nu(G)]^n [e^{t\nu}(G)], \quad t \in \mathbb{R}. \end{aligned}$$

References

- [1] L. N. Argabright and J. G. Lamadrid, *Fourier Analysis of Unbounded Measures on Locally Compact Abelian Groups*, Mem. Amer. Math. Soc. No. 145, 1974.
- [2] L. N. Argabright and J. G. Lamadrid, *Almost Periodic Measures*, Mem. Amer. Math. Soc. No. 428, 1990.
- [3] V. Barbu, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Editura Academiei, Bucureşti, 1976.
- [4] N. Dinculeanu, *Vector Measures*, Veb Deutscher Verlag Der Wissenschaften, Berlin, 1966.
- [5] N. Dinculeanu, *Integrarea pe Spaţii Local Compacte*, (in romanian), Editura Academiei R.P.R., Bucureşti, 1965.
- [6] J. A. Goldstein, *Semigroups of Linear Operators and Applications*, Oxford University Press, New York, 1985.
- [7] E. Hewitt and K. A. Ross, *Abstract Harmonic Analysis*. Vol. I. Springer - Verlag, Berlin, Göttingen, Heidelberg, 1963.
- [8] E. Hille and R. S. Phillips, *Functional Analysis and Semi - Groups*, American Mathematical Society, Providence, 1957.
- [9] J. G. Lamadrid, *Sur les Mesures Presque Périodiques*, Séminaire KGB sur les marches aléatoires, Astérisque 4, 1973.
- [10] W. Rudin, *Fourier Analysis on Groups*, Interscience Tracts in Pure and Applied Mathematics, Number 12, Interscience Publishers – John Wiley and Sons, New York, London, 1962.
- [11] T. Vladislav and I. Raşa, *Analiză Numerică. Aproximare, Problema lui Cauchy Abstractă, Proiectori Altomare*, (in romanian), Editura Tehnică, Bucureşti, 1999.

ITERATIVE PROCEDURE FOR STABILIZING SOLUTIONS OF DIFFERENTIAL RICCATI TYPE EQUATIONS ARISING IN STOCHASTIC CONTROL

Vasile Dragan and Toader Morozan
*Institute of Mathematics of the Romanian Academy,
P.O. Box 1-764,
Ro-70700, Bucharest,
Romania
vdragan@fx.ro*

Adrian M. Stoica
*University "Politehnica" of Bucharest
Faculty of Aerospace Engineering
Str. Splaiul Independentei, no. 313,
Ro-77206, Bucharest,
Romania
amstoica@fx.ro*

Abstract The paper presents a numerical algorithm to determine the stabilizing solutions of the Riccati type differential equations arising in the optimal control of time-varying stochastic systems subjected both to multiplicative white noise and to Markov jumps.

Keywords: Stochastic linear systems, optimal control, stabilizing solutions of Riccati type systems, iterative procedures

1. Riccati equations arising in stochastic control

In the present paper the following Riccati type system of differential equations is considered:

$$\begin{aligned}
& \frac{d}{dt} X(t, i) + A_0^*(t, i)X(t, i) + X(t, i)A_0(t, i) + \sum_{k=1}^r A_k^*(t, i)X(t, i)A_k(t, i) \\
& + \sum_{j=1}^d q_{ij}X(t, j) - (X(t, i)B_0(t, i) + \sum_{k=1}^r A_k^*(t, i)X(t, i)B_k(t, i)) \\
& + L(t, i)(R(t, i) + \sum_{k=1}^r B_k^*(t, i)X(t, i)B_k(t, i))^{-1}(B_0^*(t, i)X(t, i) + \\
& \sum_{k=1}^r B_k^*(t, i)X(t, i)A_k(t, i) + L^*(t, i)) + M(t, i) = 0, i \in \mathcal{D}.
\end{aligned} \tag{1}$$

This system arises in optimal control problems associated with the time-varying stochastic systems subjected both to multiplicative white noise and to Markovian jumping:

$$\begin{aligned}
dx(t) &= [A_0(t, \eta(t))x(t) + B_0(t, \eta(t))u(t)] dt \\
&+ \sum_{k=1}^r [A_k(t, \eta(t))x(t) + B_k(t, \eta(t))u(t)] dw_k(t)
\end{aligned} \tag{2}$$

where $t \in \mathbf{R}_+$, with the state vector $x \in \mathbf{R}^n$ and with the control inputs $u \in \mathbf{R}^m$, $\eta(t), t \geq 0$ is a right Markov chain with the state space $\mathcal{D} = \{1, \dots, d\}$ and the probability transition matrix $P(t) = [p_{ij}(t)] = e^{Qt}$, $t \geq 0$ in which the elements q_{ij} of Q have the property that $\sum_{i=1}^r q_{ij} = 0$, $i \in \mathcal{D}$ and $q_{ij} \geq 0$ if $i \neq j$. $w(t) = (w_1(t), \dots, w_r(t))^*$ is an r -dimensional standard Wiener process. Throughout the paper it is assumed that the coefficients of (1) are bounded and continuous matrix valued functions. By \mathcal{S}_n^d it is denoted the $\mathcal{S}_n \oplus \dots \oplus \mathcal{S}_n$, \mathcal{S}_n being the subspace of symmetric $n \times n$ matrices. For the next developments the following two definitions are reminded:

Definition 1. The system $(A_0, A_1, \dots, A_r; Q)$ is exponentially stable in mean square (ESMS) if there exist $\alpha > 0, \beta \geq 1$ such that

$$E[|\Phi(t, t_0)x_0|^2 | \eta(t_0) = i] \leq \beta e^{-\alpha(t-t_0)}|x_0|^2, \forall t \geq t_0 \geq 0, x_0 \in \mathbf{R}^n, i \in \mathcal{D},$$

where $\Phi(t, t_0)$ is the fundamental matrix of the linear differential equation obtained from (2) with $u \equiv 0$.

Definition 2. A solution $\tilde{X} : \mathbf{R}_+ \rightarrow \mathcal{S}_n^d$ of the equation (1) is called *stabilizing solution* if it has the following properties:

$$a) \inf_{t \geq 0} |\det[R(t, i) + \sum_{k=1}^r B_k^*(t, i)\tilde{X}(t, i)B_k(t, i)]| > 0, i \in \mathcal{D}.$$

b) The system $(A_0 + B_0 \tilde{F}, A_1 + B_1 \tilde{F}, \dots, A_r + B_r \tilde{F}; Q)$ is stable, where $\tilde{F}(t) = (\tilde{F}(t, 1), \tilde{F}(t, 2), \dots, \tilde{F}(t, d))$,

$$\begin{aligned}\tilde{F}(t, i) &= -[R(t, i) + \sum_{k=1}^r B_k^*(t, i) \tilde{X}(t, i) B_k(t, i)]^{-1} [B_0(t, i) \tilde{X}(t, i) \\ &\quad + \sum_{k=1}^r B_k^*(t, i) \tilde{X}(t, i) A_k(t, i) + L^*(t, i)], \quad (t, i) \in \mathbf{R}_+ \times \mathcal{D}.\end{aligned}$$

Consider the cost function:

$$\begin{aligned}J(t_0, x_0, u) &= E \int_{t_0}^{\infty} [x_u^*(t) M(t, \eta(t)) x_u(t) + x_u^*(t) L(t, \eta(t)) u(t) \\ &\quad + u^*(t) L^*(t, \eta(t)) x_u(t) + u^*(t) R(t, \eta(t)) u(t)] dt\end{aligned}\quad (3)$$

where $M(t, i) = M^*(t, i)$; $R(t, i) = R^*(t, i)$, $(t, i) \in \mathbf{R}_+ \times \mathcal{D}$ and $x_u(t)$ denotes the solution of the system (2) corresponding to the input $u(\cdot)$ with the initial condition $(t_0, x_0) \in \mathbf{R}_+ \times \mathbf{R}^n$. In [4] it is proved that the solution of the optimal control problem requiring to minimize the above cost function under the constraint $\lim_{t \rightarrow \infty} E|x_u(t, t_0, x_0)|^2 = 0$ depends on the stabilizing solution of the Riccati system (1). In the particular case when $A_k = 0, B_k = 0, k = 1, \dots, r$ and $\mathcal{D} = \{1\}$ equation (1) reduces to the well-known deterministic control Riccati equation considered in the pioneering work of Kalman [8]. If $A_k = 0, B_k = 0, k = 1, \dots, r$ and $d > 2$, one obtains the case when the system is subjected only to Markovian jumps for which the quadratic optimization problem has been studied for instance in [1], [7], [9], [10] and [11]. If $\mathcal{D} = \{1\}$ one gets the optimal control problem for stochastic systems with multiplicative white noise, previously considered in [2], [3], [5], [6], [12], [13].

2. An iterative procedure to compute the stabilizing solution of Riccati type systems

In this section an iterative procedure to compute the stabilizing solution of (1) is presented. The Riccati type system (1) can we rewritten in a compact form as:

$$\frac{d}{dt} X(t) + \mathcal{L}^*(t) X(t) - \mathcal{P}^*(t, X(t)) \mathcal{R}^{-1}(t, X(t)) \mathcal{P}(t, X(t)) + M(t) = 0 \quad (4)$$

$\mathcal{L}^*(t)$ being the adjoint of the operator \mathcal{L} defined as $\mathcal{L}(t) : \mathcal{S}_n^d \rightarrow \mathcal{S}_n^d$ by

$$\begin{aligned}(\mathcal{L}(t) X)(i) &= A_0(t, i) X(i) + X(i) A_0^*(t, i) + \\ &\quad + \sum_{k=1}^r A_k(t, i) X(i) A_k^*(t, i) + \sum_{j=1}^d q_{ji} X(j)\end{aligned}\quad (5)$$

$t \geq 0, i \in \mathcal{D}, X \in \mathcal{S}_n^d$, and \mathcal{P} , \mathcal{R} and M are defined as:

$$X \rightarrow \mathcal{P}(t, X) : \mathcal{S}_n^d \rightarrow \mathcal{M}_{m,n}^d$$

$$\begin{aligned}
\mathcal{P}(t, X) &= (\mathcal{P}_1(t, X), \mathcal{P}_2(t, X), \dots, \mathcal{P}_d(t, X)), \\
\mathcal{P}_i(t, X) &= B_0^*(t, i)X(i) + \sum_{k=1}^r B_k^*(t, i)X(i)A_k(t, i) + L^*(t, i) \\
X &\rightarrow \mathcal{R}(t, X) : \mathcal{S}_n^d \rightarrow \mathcal{S}_m^d \text{ by} \\
\mathcal{R}(t, X) &= (\mathcal{R}_1(t, X), \mathcal{R}_2(t, X), \dots, \mathcal{R}_d(t, X)), \\
\mathcal{R}_i(t, X) &= R(t, i) + \sum_{k=1}^r B_k^*(t, i)X(i)B_k(t, i), \\
M(t) &= (M(t, 1), M(t, 2), \dots, M(t, d)) \in \mathcal{S}_n^d.
\end{aligned}$$

The concept of stochastic stabilizability is introduced in standard manner (see, e.g. [4]).

Lemma 3. Assume that the system (2) is stochastically stabilizable. Let $\tilde{F}_0(t) = (\tilde{F}_0(t, 1), \tilde{F}_0(t, 2), \dots, \tilde{F}_0(t, d))$ be a stabilizing feedback gain and let $X_0(t) = (X_0(t, 1), \dots, X_0(t, d))$ be a bounded with bounded derivative solution of the linear differential inequality on \mathcal{S}_n^d :

$$\frac{d}{dt}X_0(t) + \mathcal{L}_{\tilde{F}_0}^*(t)X_0(t) + M_0(t) \leq 0 \quad (6)$$

where $M_0(t) = (M_0(t, 1), M_0(t, 2), \dots, M_0(t, d))$, $M_0(t, i) = M(t, i) + \varepsilon I_n + L(t, i)\tilde{F}_0(t, i) + \tilde{F}_0^*(t, i)L^*(t, i) + \tilde{F}_0(t, i)R(t, i)\tilde{F}_0(t, i)$, $\varepsilon > 0$ be fixed.

Under the considered assumptions,

$$X_0(t) - \hat{X}(t) \gg 0 \quad (7)$$

for any bounded solution $\hat{X}(t) : \mathbf{R}_+ \rightarrow \mathcal{S}_n^d$ of the differential inequality on \mathcal{S}_n^d :

$$\frac{d}{dt}X(t) + \mathcal{L}^*(t)X(t) - \mathcal{P}^*(t, X(t))\mathcal{R}^{-1}(t, X(t))\mathcal{P}(t, X(t)) + M(t) \gg 0 \quad (8)$$

satisfying the condition

$$\mathcal{R}_i(t, \hat{X}(t)) \geq \hat{\rho}I_n \quad (9)$$

for all $(t, i) \in \mathbf{R}_+ \times \mathcal{D}$, $\hat{\rho} > 0$ being a constant.

Proof. If $\hat{X}(t)$ is a bounded solution of (8) which verifies (9) one can define $\hat{M}(t) = (\hat{M}(t, 1), \hat{M}(t, 2), \dots, \hat{M}(t, d))$ by

$$\hat{M}(t) = \frac{d}{dt}\hat{X}(t) + \mathcal{L}^*(t)\hat{X}(t) - \mathcal{P}^*(t, \hat{X}(t))\mathcal{R}^{-1}(t, \hat{X}(t))\mathcal{P}(t, \hat{X}(t)) + M(t), \quad (10)$$

$t \in \mathbf{R}_+$. Clearly $\hat{M}(t) \geq 0$, by direct computations it follows that that

$$\begin{aligned} & \frac{d}{dt} \hat{X}(t) + \mathcal{L}_{\tilde{F}_0}^*(t) \hat{X}(t) + M(t) + L(t) \tilde{F}_0(t) + \tilde{F}_0^*(t) L(t) + \tilde{F}_0^*(t) R(t) \tilde{F}_0(t) \\ & - \hat{M}(t) - (\hat{F}(t) - \tilde{F}_0(t))^* \mathcal{R}(t, \hat{X}(t)) (\hat{F}(t) - \tilde{F}_0(t)) = 0 \end{aligned} \quad (11)$$

where $\hat{F}(t) = (\hat{F}(t, 1), \hat{F}(t, 2), \dots, \hat{F}(t, d))$ with

$$\hat{F}(t, i) = -\mathcal{R}_i^{-1}(t, \hat{X}(t)) \mathcal{P}_i(t, \hat{X}(t)), \quad t \in \mathcal{I}, \quad i \in \mathcal{D}. \quad (12)$$

From (11) and (6) one gets:

$$\begin{aligned} & \frac{d}{dt} (X_0(t) - \hat{X}(t)) + \mathcal{L}_{\tilde{F}_0}^*(t) (X_0(t) - \hat{X}(t)) + (\tilde{F}_0(t) - \hat{F}(t))^* \mathcal{R}(t, \hat{X}(t)) \\ & (\tilde{F}_0(t) - \hat{F}(t)) + \varepsilon J^d + \hat{M}(t) \leq 0, \quad t \geq 0. \end{aligned}$$

from which it follows that $X_0(t) - \hat{X}(t) \geq Y(t)$ where $t \rightarrow Y(t) = (Y(t, 1), Y(t, 2), \dots, Y(t, d))$ is the unique bounded solution of the Lyapunov type equation

$$\frac{d}{dt} Y(t) + \mathcal{L}_{\tilde{F}_0}^*(t) Y(t) + \varepsilon J^d = 0. \quad (13)$$

Let $T_0(t, s)$ be the linear evolution operator on \mathcal{S}_n^d , defined by the linear differential equation:

$$\frac{d}{dt} S(t) = \mathcal{L}_{\tilde{F}_0}(t) S(t).$$

Since $\tilde{F}_0(t)$ is a stabilizing feedback gain, then there exist positive constants β_0, α_0 such that $\|T_0(t, s)\| \leq \beta_0 e^{-\alpha_0(t-s)}$, $(\forall) t \geq s \geq 0$. Therefore the unique bounded solution of the equation (13) is uniform positive and the proof is complete. Based on (7) one deduces that there exists $\mu_0 > 0$, such that $\mathcal{R}_i(t, X_0(t)) \geq \mu_0 I_n$, $t \in \mathcal{I}, i \in \mathcal{D}$. Hence the feedback gain $F_0(t) = (F_0(t, 1), \dots, F_0(t, d))$ is well defined by

$$F_0(t, i) = -\mathcal{R}_i^{-1}(t, X_0(t)) \mathcal{P}_i(t, X_0(t)), \quad i \in \mathcal{D}, \quad t \in \mathcal{I}. \quad (14)$$

Further one will prove that $F_0(t)$ is a stabilizing feedback gain for the system (2). To this end one considers $\hat{X}(t)$ a bounded solution of (8) which verifies condition (9). By direct computation and using (10) and (14) one obtains:

$$\begin{aligned} & \frac{d}{dt} \hat{X}(t) + \mathcal{L}_{F_0}^*(t) \hat{X}(t) + M(t) + L(t) F_0(t) + F_0^*(t) L^*(t) + \\ & F_0^*(t) R(t) F_0(t) - (\hat{F}(t) - F_0(t))^* \mathcal{R}(t, \hat{X}(t)) (\hat{F}(t) - F_0(t)) - \hat{M}(t) = 0. \end{aligned} \quad (15)$$

Further (6) may be rewritten as:

$$\begin{aligned} \frac{d}{dt}X_0(t) + \mathcal{L}_{F_0}^*(t)X_0(t) + M(t) + L(t)F_0(t) + F_0^*(t)L^*(t) + F_0^*(t)R(t)F_0(t) \\ + (F_0(t) - \tilde{F}_0(t))^* \mathcal{R}(t, X_0(t))(F_0(t) - \tilde{F}_0(t)) + \varepsilon J^d \leq 0. \end{aligned} \quad (16)$$

From (15), (16) and (7) it follows that $t \rightarrow X_0(t) - \hat{X}(t)$ is a bounded and uniform positive solution of the linear differential inequality on \mathcal{S}_n^d :

$$\frac{d}{dt}X(t) + \mathcal{L}_{F_0}^*(t)X(t) + \frac{\varepsilon}{2}J^d \ll 0.$$

and then, since $X_0(t) - \hat{X}(t) \gg 0$ it results that the system $(A_0 + B_0F_0, A_1 + B_1F_0, \dots, A_r + B_rF_0; Q)$ is stable which shows that $F_0(t)$ is a stabilizing feedback gain. As a consequence we deduce that for each $i \in \mathcal{D}$, the zero state equilibrium of the linear differential equation on \mathbf{R}^n ,

$$\frac{d}{dt}X(t) = (A_0(t, i) + \frac{1}{2}q_{ii}I_n + B_0(t, i)F_0(t, i))X(t)$$

is exponentially stable. Particularly in the time invariant case it follows that the eigenvectors of the matrices $A_0(i) + \frac{1}{2}q_{ii}I_n + B_0(i)F_0(i)$ are located in the half plane $Re\lambda < 0$.

Taking $X_0(t), F_0(t)$ as a first step, one constructs iteratively the sequences: $\{X_l(t, i)\}_{l \geq 0}$, $\{F_l(t, i)\}_{l \geq 0}, i \in \mathcal{D}$ as follows: $t \rightarrow X_{l+1}(t, i)$ is the unique bounded solution of the Lyapunov equation

$$\begin{aligned} \frac{d}{dt}X_{l+1}(t, i) + [\tilde{A}_0(t, i) + B_0(t, i)F_l(t, i)]^*X_{l+1}(t, i) \\ + X_{l+1}(t, i)[\tilde{A}_0(t, i) + B_0(t, i)F_l(t, i)] + M_{l+1}(t, i) = 0 \end{aligned} \quad (17)$$

where $M_{l+1}(t) = (M_{l+1}(t, 1) \dots M_{l+1}(t, d))$ with

$$\begin{aligned} M_{l+1}(t, i) &= M(t, i) + \frac{\varepsilon}{l+2}I_n + L(t, i)F_l(t, i) \\ &\quad + F_l^*(t, i)L^*(t, i) + F_l^*(t, i)R(t, i)F_l(t, i) \\ &\quad + \sum_{k=1}^r [A_k(t, i) + B_k(t, i)F_l(t, i)]^*X_l(t, i) \\ &\quad \times [A_k(t, i) + B_k(t, i)F_l(t, i)] + \sum_{j \neq i} q_{ij}X_l(t, j) \end{aligned} \quad (18)$$

$$\begin{aligned} \tilde{A}_0(t, i) &= A_0(t, i) + \frac{1}{2}q_{ii}I_n \\ F_{l+1}(t, i) &= -(R(t, i) + \sum_{k=1}^r B_k^*(t, i)X_l(t, i)B_k(t, i))^{-1}(B_0^*(t, i)X_{l+1}(t, i)) \end{aligned}$$

$$+ \sum_{k=1}^r B_k^*(t, i) X_l(t, i) A_k(t, i) + L^*(t, i)), \quad l \geq 0, i \in \mathcal{D}.$$

Further one proves that

- a) $X_l(t, i) - \hat{X}(t, i) \geq \mu_l I_n > 0$ for all integers $l \geq 0$, $i \in \mathcal{D}$, $t \in \mathcal{I}$, $\hat{X}(t) = (\hat{X}(t, 1) \dots \hat{X}(t, d))$ being an arbitrary bounded solution of (8) which verifies (9) and μ_l is a positive constant which not depend upon $\hat{X}(t)$.
- b) The zero state equilibrium of the linear differential equation on \mathbf{R}^n

$$\frac{d}{dt} x(t) = [\tilde{A}_0(t, i) + B_0(t, i) F_l(t, i)] x(t)$$

is exponentially stable for each $i \in \mathcal{D}, l \geq 0$.

- c) $X_l(t, i) \geq X_{l+1}(t, i), \forall l \geq 0, (t, i) \in \mathcal{I} \times \mathcal{D}$.

Notice that the properties a) and b) have been proved for $l = 0$. It will be shown by induction that a), b), c) are fulfilled for every $l \geq 0$. To this end assume that a), b), c) are fulfilled for the first $l - 1$ terms of the sequences defined by (17) and (18). By direct computation one obtains that if $\hat{X}(t)$ is a bounded solution of the inequality (8) which verifies (9) then

$$\begin{aligned} & \frac{d}{dt} \hat{X}(t, i) + [\tilde{A}_0(t, i) + B_0(t, i) F_{l-1}(t, i)]^* \hat{X}(t, i) \\ & + \hat{X}(t, i) (\tilde{A}_0(t, i) + B_0(t, i) F_{l-1}(t, i)) \\ & + \sum_{k=1}^r [A_k(t, i) + B_k(t, i) F_{l-1}(t, i)]^* \hat{X}(t, i) [A_k(t, i) + B_k(t, i) F_{l-1}(t, i)] \\ & + \sum_{j=1, j \neq i}^d q_{ij} \hat{X}(t, j) + M(t, i) + L(t, i) F_{l-1}(t, i) + F_{l-1}^*(t, i) L^*(t, i) \\ & + F_{l-1}^*(t, i) R(t, i) F_{l-1}(t, i) - \hat{M}(t, i) - (\hat{F}(t, i) - F_{l-1}(t, i))^* \mathcal{R}_i(t, \hat{X}(t)) \\ & \times (\hat{F}(t, i) - F_{l-1}(t, i)) = 0, \end{aligned} \tag{19}$$

$\hat{M}(t, i), \hat{F}(t, i)$ being defined in (10) and (12) respectively.

Using (17) with l replaced by $l - 1$ one gets

$$\begin{aligned} & \frac{d}{dt} [X_l(t, i) - \hat{X}(t, i)] + [\tilde{A}_0(t, i) + B_0(t, i) F_{l-1}(t, i)]^* [X_l(t, i) - \hat{X}(t, i)] \\ & + [X_l(t, i) - \hat{X}(t, i)] [\tilde{A}_0(t, i) + B_0(t, i) F_{l-1}(t, i)] + \frac{\varepsilon}{l+1} I_n + \Delta_l(t, i) = 0 \end{aligned} \tag{20}$$

where

$$\begin{aligned} \Delta_l(t, i) = & \sum_{k=1}^r [A_k(t, i) + B_k(t, i) F_{l-1}(t, i)]^* [X_{l-1}(t, i) - \hat{X}(t, i)] \\ & \times [A_k(t, i) + B_k(t, i) F_{l-1}(t, i)] + \sum_{j=1, j \neq i}^d q_{ij} (X_{l-1}(t, j) - \hat{X}(t, j)) \\ & + \hat{M}(t, i) + (\hat{F}(t, i) - F_{l-1}(t, i))^* \mathcal{R}_i(t, \hat{X}(t)) (\hat{F}(t, i) - F_{l-1}(t, i)). \end{aligned}$$

Since $X_{l-1}(t, i) - \hat{X}(t, i) \geq \mu_{l-1}I_n$ we get $\Delta_l(t, i) \geq 0$. Taking into account that $\tilde{A}_0(t, i) + B_0(t, i)F_l(t, i)$ generates an exponentially stable evolution we may conclude that the equation (20) has a unique bounded solution which is uniform positive definite. Hence there exists $\mu_l > 0$, such that $X_l(t, i) - \hat{X}(t, i) \geq \mu_l I_n$ and thus a) is fulfilled. Further it follows that $\mathcal{R}_i(t, X_l(t)) \geq \nu_l I_m > 0$.

Using (18) one may write

$$\begin{aligned} \frac{d}{dt}X_l(t, i) + [\tilde{A}_0(t, i) + B_0(t, i)F_l(t, i)]^*X_l(t, i) + X_l(t, i)[\tilde{A}_0(t, i) \\ + B_0(t, i)F_l(t, i)] + \sum_{k=1}^r[A_k(t, i) + B_k(t, i)F_l(t, i)]^*X_{l-1}(t, i)[A_k(t, i) \\ + B_k(t, i)F_l(t, i)] + \sum_{j=1, j \neq i}^d q_{ij}X_{l-1}(t, j) + M(t, i) + \frac{\varepsilon}{l+1}I_n \\ + L(t, i)F_l(t, i) + F_l^*(t, i)L^*(t, i) + F_l^*(t, i)R(t, i)F_l(t, i) + [F_l(t, i) \\ - F_{l-1}(t, i)]^*\mathcal{R}_i(t, X_{l-1}(t))[F_l(t, i) - F_{l-1}(t, i)] = 0. \end{aligned} \quad (21)$$

It is easy to see that $t \rightarrow \hat{X}(t, i)$ verifies:

$$\begin{aligned} \frac{d}{dt}\hat{X}(t, i) + (\tilde{A}_0(t, i) + B_0(t, i)F_l(t, i))^*\hat{X}(t, i) + \hat{X}(t, i)(\tilde{A}_0(t, i) \\ + B_0(t, i)F_l(t, i)) + \sum_{k=1}^r(A_k(t, i) + B_k(t, i)F_l(t, i))^*\hat{X}(t, i)(A_k(t, i) \\ + B_k(t, i)F_l(t, i)) + \sum_{j=1, j \neq i}^d q_{ij}\hat{X}(t, j) + M(t, i) + F_l^*(t, i)L^*(t, i) \\ + L(t, i)F_l(t, i) + F_l^*(t, i)R(t, i)F_l(t, i) - \hat{M}(t, i) \\ - (\hat{F}(t, i) - F_l(t, i))\mathcal{R}_i(t, \hat{X}(t))(\hat{F}(t, i) - F_l(t, i)) = 0. \end{aligned} \quad (22)$$

Thus it results that for each $i \in \mathcal{D}$, $t \rightarrow X_l(t, i) - \hat{X}(t, i)$ is a bounded and uniformly positive definite solution of the linear differential inequality:

$$\begin{aligned} \frac{d}{dt}Y(t, i) + [\tilde{A}_0(t, i) + B_0(t, i)F_l(t, i)]^*Y(t, i) \\ + Y(t, i)[\tilde{A}_0(t, i) + B_0(t, i)F_l(t, i)] + \frac{\varepsilon}{2(l+1)}I_n < 0 \end{aligned}$$

which allows to conclude that the zero state equilibrium of the linear differential equation

$$\frac{d}{dt}x(t) = (\tilde{A}_0(t, i) + B_0(t, i)F_l(t, i))x(t) \quad (23)$$

is exponentially stable and b) is fulfilled.

Subtracting (17) from (21) one gets that $t \rightarrow X_l(t, i) - X_{l+1}(t, i)$ is a bounded solution of the equation

$$\begin{aligned} \frac{d}{dt}(X_l(t, i) - X_{l+1}(t, i)) + (\tilde{A}_0(t, i) + B_0(t, i)F_l(t, i))^*(X_l(t, i) - X_{l+1}(t, i)) \\ + (X_l(t, i) - X_{l+1}(t, i))(\tilde{A}_0(t, i) + B_0(t, i)F_l(t, i)) + \hat{\Delta}_l(t, i) = 0 \end{aligned} \quad (24)$$

where

$$\begin{aligned}\hat{\Delta}_l(t, i) = & \frac{\varepsilon}{(l+1)(l+2)} I_n + [F_l(t, i) - F_{l-1}(t, i)]^* \mathcal{R}_i(t, X_{l-1}(t)) \\ & \times [F_l(t, i) - F_{l-1}(t, i)] + \sum_{k=1}^r [A_k(t, i) + B_k(t, i) F_l(t, i)]^* \\ & \times (X_{l-1}(t, i) - X_l(t, i)) [A_k(t, i) + B_k(t, i) F_l(t, i)] \\ & + \sum_{j=1, j \neq i}^d q_{ij} (X_{l-1}(t, j) - X_l(t, j)),\end{aligned}\quad (25)$$

for $l \geq 1$ and

$$\hat{\Delta}_l(t, i) \geq \frac{\varepsilon}{2} I_n + (F_0(t, i) - \tilde{F}_0(t, i))^* \mathcal{R}_i(t, X_0(t)) (F_0(t, i) - \tilde{F}_0(t, i))$$

for $l = 0$.

Since $\hat{\Delta}_0(t, i) \geq 0$ and the zero state equilibrium of (23) for $l = 0$ is exponentially stable it follows from (24) for $l = 0$ that $X_0(t, i) - X_1(t, i) \geq 0$ and further by induction one obtains that $\hat{\Delta}_l \geq 0$ for $l \geq 1$ which leads to $X_l(t, i) - X_{l+1}(t, i) \geq 0$ and c) is fulfilled.

From a) and c) one concludes that the sequences $\{X_l(t, i)\}_{l \geq 0}, i \in \mathcal{D}$ are convergent. More precisely we have:

Theorem 1 Assume that system $(\mathbf{A}, \mathbf{B}; Q)$ is stochastically stabilizable and that the differential inequality (8) has a bounded with bounded derivative solution $\hat{X}(t)$ which satisfies (9). Then for any choice of a stabilizing feedback gain $\tilde{F}_0(t) = (\tilde{F}_0(t, 1), \tilde{F}_0(t, 2), \dots \tilde{F}_0(t, d))$, the sequences $\{X_l(t, i)\}_{l \geq 0}, i \in \mathcal{D}$, constructed as solutions of (17) (the first terms $X_0(t, i)$ obtained by solving (6)) are convergent. If

$$\tilde{X}(t, i) = \lim_{l \rightarrow \infty} X_l(t, i), \quad (t, i) \in \mathcal{I} \times \mathcal{D} \quad (26)$$

then $\tilde{X}(t) = (\tilde{X}(t, 1), \tilde{X}(t, 2) \dots \tilde{X}(t, d))$ is the stabilizing bounded solution of the system (1) verifying (9).

Remark 1. a) Excepting the first step, when to obtain $X_0(t, i)$ one needs to solve a system of linear inequalities of higher dimension, namely (6), further to obtain the next terms of the sequences $\{X_l(t, i)\}_{l \geq 1}, i \in \mathcal{D}$, one needs to solve a system of d uncoupled Lyapunov equations. Note that to compute the gains $F_l(t, i)$ in (18) we need both the value of $X_l(t, i)$ and the value of $X_{l-1}(t, i)$.

b) Based on the uniqueness of the bounded solution of a Lyapunov equation, it follows that when the system (2) and the cost function (3) are in the time invariant case, then the matrices X_l and F_l do not depend upon t and they are obtained as solutions of algebraic Lyapunov equations.

c) Also from the uniqueness of the bounded solution of a Lyapunov equation we deduce that if the coefficients of the system (1) are θ -periodic functions defined on \mathbf{R} , then the bounded solution of (17) are

θ -periodic function too. Hence it is sufficient to compute the values of $X_l(t, i), F_l(t, i)$ on the interval $[0, \theta]$. At each step l , the initial condition $X_l(0, i)$ is obtained by solving the linear equation

$$X_l(0, i) = \Phi_{l,i}^*(\theta, 0)X_l(0, i)\Phi_{l,i}(\theta, 0) + \int_0^\theta \Phi_{l,i}^*(s, 0)M_l(s, i)\Phi_{l,i}(s, 0)ds$$

$\Phi_{l,i}(t, s)$ being the fundamental matrix solution of (23). For the first step $X_0(t, i)$ is chosen as a periodic solution of the Lyapunov type equation on \mathcal{S}_n

$$\frac{d}{dt}X_0(t) + \mathcal{L}_{\tilde{F}_0}^*(t)X_0(t) + M_0(t) = 0$$

where $M_0(t) = (M_0(t, 1), M_0(t, 2), \dots M_0(t, d))$,

$$\begin{aligned} M_0(t, i) &= M(t, i) + \varepsilon I_n + L(t, i)\tilde{F}_0(t, i) + \tilde{F}_0^*(t, i)L^*(t, i) \\ &\quad + \tilde{F}_0^*(t, i)R(t, i)\tilde{F}_0(t, i). \end{aligned}$$

If $T_0(t, t_0)$ is the linear evolution operator defined by the linear differential equation on \mathcal{S}_n^d :

$$\frac{d}{dt}S(t) = \mathcal{L}_{\tilde{F}_0}(t)S(t) \quad (27)$$

then the initial condition $X_0(0) = (X_0(0, 1), X_0(0, 2), \dots X_0(0, d))$ is given by

$$X_0(0) = [\tilde{J} - T_0^*(\theta, 0)]^{-1} \int_0^\theta T_0^*(s, 0)M_0(s)ds$$

where \tilde{J} is the identity operator on \mathcal{S}_n^d ; $\tilde{J} - T_0^*(\theta, 0)$ is invertible due to the exponential stability of the evolution defined by the differential equation (27).

3. Numerical examples

The above iterative numerical procedures will be illustrated considering the linear time-invariant stochastic system of order $n = 2$, subjected to both multiplicative noise and Markovian jumps with $r = 1$ and $\mathcal{D} = \{1, 2\}$ having:

$$\begin{aligned} A_0(1) &= \begin{bmatrix} -1 & 0 \\ 1 & -1 \end{bmatrix}, & A_0(2) &= \begin{bmatrix} -1 & 1 \\ 0 & -1 \end{bmatrix}, \\ A_1(1) &= \begin{bmatrix} -1 & 1 \\ 0 & -2 \end{bmatrix}, & A_1(2) &= \begin{bmatrix} -2 & 1 \\ 1 & -1 \end{bmatrix}, \end{aligned}$$

$$\begin{aligned} B_0(1) &= \begin{bmatrix} 1 \\ -1 \end{bmatrix}, & B_0(2) &= \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \\ L_0(1) &= \begin{bmatrix} 1 \\ -1 \end{bmatrix}, & L_0(2) &= \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \\ M_0(1) &= \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}, & M_0(2) &= \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}, \\ R(1) &= 1, & R(2) &= 2. \end{aligned}$$

The purpose is to solve the system (1) correspondind to the above numerical values using the iterative procedure indicated in the statement of Theorem 1. Three dinstinct cases have been considered: the case when the system is subjected only to Markov jumps, the case when the system is subjected only to multiplicative white noise, and the case when the system is perturbed with both multiplicative white noise and Markovian jumps.

Case a). The Markovian jumping case:

$A_1(i) = 0$, $B_1(i) = 0$, $i \in \mathcal{D}$. For the stabilizing gains

$$\tilde{F}_0(1) = [0.5923 \quad -0.7004], \quad \tilde{F}_0(2) = [-0.0330 \quad 0.0653],$$

solving (6) one obtains:

$$X_0(1) = 10^3 \begin{bmatrix} 1.5519 & -0.0524 \\ -0.0524 & 1.7776 \end{bmatrix}, \quad X_0(2) = 10^3 \begin{bmatrix} 1.1139 & 0.2680 \\ 0.2680 & 1.3970 \end{bmatrix}.$$

The solution of (1) for this case was determined solving iteratively (17). For an imposed level of accuracy $\|X_{l+1}(i) - X_l(i)\| < 10^{-6}$ we obtained after 69 iterations:

$$X(1) = \begin{bmatrix} 30.7868 & 24.3960 \\ 24.3960 & 26.2218 \end{bmatrix}, \quad X(2) = \begin{bmatrix} 21.5504 & -11.7226 \\ -11.7226 & 19.2254 \end{bmatrix}.$$

Case b). The multiplicative white noise perturbations case: $\mathcal{D} = \{1\}$, $A_i = A_i(1)$; $B_i = B_i(1)$, $i = 0, 1$.

In this case one obtains the initial values:

$$\tilde{F}_0 = [-0.4094 \quad 0.8482], \quad X_0 = \begin{bmatrix} 292.8945 & 163.9337 \\ 163.9337 & 140.9240 \end{bmatrix}$$

and, after 202 iterations, the solution of (1):

$$X = \begin{bmatrix} 1.0782 & 1.0307 \\ 1.0307 & 0.5878 \end{bmatrix}.$$

Case c). The case when the system is subjected to both Markovian jumps and multiplicative white noise.

In this situation the following initial values are obtained:

$$\begin{aligned}\tilde{F}_0(1) &= [-0.3852 \quad 0.8594], \quad \tilde{F}_0(2) = [-0.9000 \quad 0.5763], \\ X_0(1) &= 10^8 \begin{bmatrix} 5.8005 & -4.5733 \\ -4.5733 & -3.7733 \end{bmatrix}, \\ X_0(2) &= 10^8 \begin{bmatrix} -0.7123 & -0.5110 \\ 0.5110 & -4.8453 \end{bmatrix}.\end{aligned}$$

The solution of (1) was obtained after 133 iterations solving (17); thus it results:

$$X(1) = \begin{bmatrix} 2.1893 & 2.0159 \\ 2.0159 & 2.0998 \end{bmatrix}, \quad X(2) = \begin{bmatrix} 0.7940 & -0.4088 \\ -0.4088 & 3.3714 \end{bmatrix}.$$

References

- [1] H. Abou-Kandil, G. Freiling, G.Jank, Solution and asymptotic behaviour of coupled Riccati equations in jump linear systems, *IEEE, Trans. Auto. Control*, **39**, 1631-1636, 1994.
- [2] J.M.Bismut, Linear quadratic optimal control with random coefficients *SIAM J. Contr. Optimiz.*, **14**, 419-444, 1976.
- [3] G. DaPrato, A. Ichikawa, Quadratic control for linear time varying systems, *SIAM J. Control Optimization*, **28**, 359-381, 1990.
- [4] V. Dragan, T. Morozan, The linear quadratic optimization problem and tracking problem for a class of linear stochastic systems with multiplicativewhite noise and Markovian jumping. Preprint IMAR nr. 3/2001.
- [5] U.G.Hausmann - Optimal stationary control with state and control dependent noise, *SIAM J. Control and Optimization*, **9**, 184-198, 1971.
- [6] A.Ichikawa, Optimal control of a linear stochastic evolution equation with state and control dependent noise, *Proc. IMA. Conference "Recent theoretical developments in control"*, Leicester, England, Academic Press, 1976.
- [7] Y. Ji, H.J.Chizeck, Controllability, stabilizability, and continuous-time Markovian jump linear quadratic control- *IEEE Trans. Auto. Control*, **35**, no.7, 777-788, 1990.
- [8] R. Kalman, Contributions on the theory of optimal control, *Buletin de la Sociedad Math. Mexicana*, Segunda Serie, **5**, 1, 102-19, 1960.
- [9] T. Morozan, Optimal stationary control for dynamic systems with Markov perturbations, *Stochastic Analysis and Applications*, **1** , 3, 299-323, 1983.
- [10] T. Morozan, Stability and control for linear system with jump Markov perturbation, *Stochastic Analysis and Applications*, **13** , 1, 91-110, 1995.
- [11] T. Morozan, Linear quadratic control and tracking problems for time-varying stochastic differential systems perturbed by a Markov chain, *Revue Roum. Math. Pure et Applique*, to appear.
- [12] Tessitore G., Some remarks on the Riccati equation arising in an optimal control problem with state and control dependent noise, *SIAM Control and Optimization*, **30**, 3, 717-744, 1992.
- [13] W.H. Wonham, Random differential equations in control theory, *Probabilistic Methods in Applied Math.*, **2**, (A.T. Barucha-Reid, Ed.) Academic Press, New-York, 131-212, 1970.

RECURSIVE DECONVOLUTION: AN OVERVIEW OF SOME RECENT RESULTS

F. Fagnani, L. Pandolfi*

Politecnico di Torino, Dip. di Matematica

fagnani@calvino.polito.it, Lucipan@polito.it

Abstract We present some recent results on recursive solution of Volterra integral equations of first kind. The method which is used is suggested by control and game theory.

Keywords: Volterra integral equations, deconvolution, Abel equation.

1. Preliminaries

The deconvolution problem has a long history and many different aspects. We are concerned here with the deconvolution problem for *causal* systems, which in the time invariant case is the solution of

$$y = k * u$$

in the unknown u . Here y , k and u are defined for $t \geq 0$ and the *kernel* k is in general a distribution supported on $t \geq 0$.

Our study of the deconvolution problem uses ideas that arose in control theory. In this context, the problem is as follows: a linear system is given,

$$\dot{x} = Ax + Bu, \quad y = Cx + Du, \quad x(0) = 0. \quad (1)$$

We want to know whether a measured output y is produced by a unique input u . If this is the case than the system is called *left invertible* or *ideally observable* in the russian literature.

If the system is left invertible then we want to construct u as the output of a new system

$$\dot{\xi} = \tilde{A}\xi + \tilde{B}y, \quad u = \tilde{C}\xi + \tilde{D}y. \quad (2)$$

*Paper supported by the Italian MURST

System (2) is the *inverse system* to (1).

We observe that the output y to system (1) is given by

$$y = k * u, \quad K(t) = D\delta + Ce^{At}B$$

(δ denotes Dirac's delta). If D is invertible and the system is time invariant, then the construction of the inverse system is trivial, $\tilde{A} = A - BD^{-1}C$, $\tilde{B} = BD^{-1}$, $\tilde{C} = -D^{-1}C$, $\tilde{D} = D^{-1}$. If D is not invertible, and in the most important case $D = 0$, the construction of the inverse system leads to an ill posed problem and its study was most fruitful, since it led to the construction of the "geometric" theory of linear finite dimensional systems in the papers and books of Basile, Marro, Morse and Wonham and, stimulated by the ideas of Krasovski in [7], the construction of an iterative scheme for the approximate inversion by Osipov [8]. We present an overview of our results, obtained in the line of the book [8].

We note that the geometric theory concerns mostly finite dimensional systems. The examples in [14] shows that extensions of the geometric theory to distributed systems can only produce weak results.

We note that in many applications the output is read only on a finite time interval $[0, T]$, at discrete times $\tau_k = k\tau$, $\tau = T/N$ and the measures are corrupted by errors of known tolerance. Hence, available data are $\xi_k = y(k\tau) + \theta_k$, $|\theta_k| < h$.

We noted that, when $D = 0$, the problem is ill posed. Hence we rely on a penalization approach for the approximate inversion of the system. The penalization approach is performed at each step and, in fact, in the overall it leads to a "shift of the spectrum" of an operator. The method depends on the introduction of an additional parameter α and constructs functions

$$v = v_{\tau, \{\xi\}, \alpha}$$

where $\{\xi\}$ is the vector of the measures. We want: 1) at time t , $v(t)$ only depends on the measures taken at $\tau_k \leq t$; 2) when τ , h and α converge to zero, while respecting suitable relations, v should converge to u in a suitable topology.

We observe that we cannot hope that v converges to u if τ , h and α converge to zero independently, since the problem is ill posed.

2. The key idea

Finite dimensional systems, both linear and non linear but with $C = I$ and $D = 0$, full state observations (and special cases of $C \neq I$), have been investigated in [8]. The key idea is to associate a "model" to the

system:

$$\dot{w} = Aw + Bv, \quad z = Cw. \quad (3)$$

and to choose v so to force z to track y . Hopefully, under suitable conditions, v will track u . A general analysis of system (1) in finite dimensional spaces and $C \neq I$ is in [2]. The proposed algorithm is as follows: the input v is piecewise continuous, updated at each step τ_k , $v(t) = v_k(t)$, $t \in [\tau_k, \tau_{k+1})$, defined by

$$v_k = \arg \min \left\{ \|Cw(\tau_{k+1}) - \xi_k\|^2 + \alpha \int_{\tau_k}^{\tau_{k+1}} \|v(s)\|^2 ds \right\}.$$

This same idea will be used to study also the case of Volterra integral equations.

We quote the papers [9, 10] for distributed systems in state space form.

3. Finite dimensional results

The class of left invertible finite dimensional systems (with $D = 0$) is geometrically characterized as those systems whose *maximal controllability subspace in $\ker C$* , denoted \mathcal{R}_* , is $\{0\}$.

The idea that we describe in sect. 2 was thoroughly analyzed in [2]. We proved:

Theorem 1 *If $\ker CB = \{0\}$ then $\lim_{\alpha \rightarrow 0} \left[\lim_{\tau \rightarrow 0, h \rightarrow 0} v_{\tau, \{\xi\}, \alpha} \right] = u$. More precisely we have the following result (where $v = v_{\tau, \{\xi\}, \alpha}$). 1) If the unknown input u is square integrable on $[0, T]$ then v converges to u in $L^2(0, T)$; 2) if the unknown input u is of class $W^{1,2}(0, T)$ then v converges to u uniformly on $[\sigma, T]$ for every $\sigma > 0$, and it converges uniformly to u on $[0, T]$ if, furthermore, $u(0) = 0$.*

A system which satisfies condition $\ker CB = \{0\}$ is a “system of relative degree 1”.

The class of left invertible systems is *larger* than the class of those systems for which $\ker CB = \{0\}$. The key instrument for the extension of the above result from the special case $\ker CB = \{0\}$ to the general case $\mathcal{R}_* = \{0\}$ is *Morse canonical form*, see [12]. We do not describe this complicated instrument here. We simply note that the use of this form reduce the problem to the case that the system is a chain of integrators, essentially a system of n_i -th order scalar equations, to which the method above can be applied step by step, as in the next example.

Example 2 Let

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = u, \quad y = x_1.$$

In this case $CB = 0$ and Theorem 1 cannot be applied. But, we can associate $w_1 = \dot{x}_2$ to the first component $\dot{x}_1 = x_2$. We observe that $x_2 \in W^{1,2}$ and $x_2(0) = 0$ so that we can give a uniform estimate $\hat{x}_2(t)$ of $x_2(t)$. At time τ_k we can apply the procedure outlined above to the system $\dot{x}_2 = u$ and we can recursively identify the input u . ■

3.1. An application

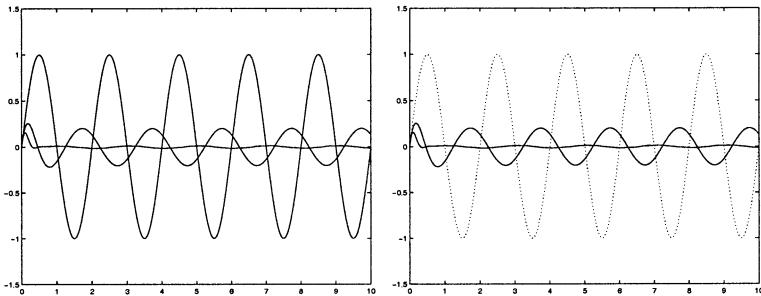
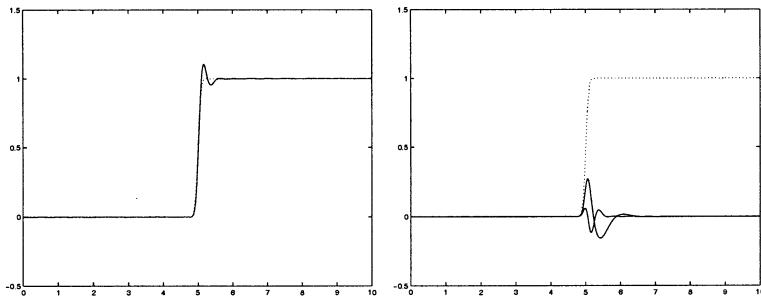
The results above have been applied to the reduction of the effect of a disturbance (internally or externally generated), in [1]. We present some simulations taken from this paper concerning the control of a robot motor, against (internally generated) disturbances, due to variations in the transported load.

The block diagram of a robot motor can be found in [6]. It is composed of two blocks: the main block is the motor itself whose output is the track (let it be X) followed by the robot. This is fed-back to an “acceleration controller” which also accept as the input the signal X^{cmd} , which is the track to be followed. It can easily be shown that the difference between the real path and the nominal path is

$$X - X^{cmd} = \frac{1}{M_0} \cdot \frac{1}{s^2 + K_1 s + K_2} v$$

where K_1 and K_2 are constants which enter in the definition of the “acceleration control”. These constants are chosen in such a way to have an asymptotically stable system. The motor and its “acceleration controller” are designed on the basis of the choice of the nominal mass M_0 that the robot should carry. In this way, if the disturbance v is equal to 0, path tracking is achieved: the values of the path X (and of its first and second derivatives) coincide with the one of X^{cmd} . Errors on the initial condition, or impulsive disturbances, generate fast transients. A persistent disturbance v however is not canceled by the “acceleration controller”.

We use the deconvolution ideas in order to *identify* and then to *cancel* the persistent disturbance due to changing loads to be transported. The results are in the plots below. The plots represent on the left X^{cmd} and X , when we apply our algorithm for disturbance reduction and, on the left, the errors $X^{cmd} - X$ with and without compensation. in two extreme cases: the case that X^{cmd} is a sinusoidal path ($\sin \pi t$) and the case that the path has an abrupt change.

Figure 1. Motor. $\tau = 0.01$, $\alpha = 1/85$, $h = 0.1$.Figure 2. Motor. $\tau = 0.01$, $\alpha = 1/85$, $h = 0.1$.

4. Distributed systems

Applications to distributed systems of the previous ideas have been widely investigated by the Ekaterinburg school. An overview is in [13], see also the paper by Maksimov in these proceedings. As the geometric theory of linear systems seems not be extendible to distributed systems, the state space analysis at the moment gives weak results, which require full state observation, see [9, 10, 15], unless the system has a known special structure. See also the case of systems with delays examined in [11]. A special case of distributed input–output system is examined in [4], in the context of degenerate systems.

It is well known that the general deconvolution problem for distributed systems (in particular for the heat equation) is a very hard problem, due to the fact that the kernel may have a zero for $t \rightarrow 0+$, of infinite order. However, this does not happen in important cases. Even more,

the kernel may be singular, as in the Abel equation

$$y(t) = \int_0^t \frac{1}{(t-s)^\gamma} u(s) \, ds, \quad 0 \leq \gamma < 1$$

which is encountered, for example, in some input–output problem of heat transmission.

The solution of an Abel equation is a classical subject, see [5] and close formulas for the solution exist, which however require the computation of the derivative of an integral (fractional derivative). In spite of the fact that close formulas are always important, numerically these formulas contains redundancies since a part of the computed derivative is killed by the presence of the integral. For this reason in the next section we describe the results that we can obtain when applying the method outlined above to a class of integral equations which includes Abel equations.

We shall distinguish the case of convolution equations, i.e. the case that the kernel depends on the difference $t - s$ from the general case since, in the convolutional case, we can use powerful frequency domain techniques.

5. Volterra integral equations

We consider a Volterra integral equation

$$y(t) = \int_0^t K(t,s)u(s) \, ds \quad t \in [0, T]$$

and we want to solve for u , on the basis of observations taken on y at the time instants $\tau_k = kT/N$.

We represent

$$y(\tau_{k+1}) = y(\tau_k) + \int_0^{\tau} K(\tau_{k+1}, \tau_k + s)u(\tau_k + s) \, ds + \int_0^{\tau_k} F(\tau_k, s)u(s) \, ds, \quad (4)$$

where

$$F(t,s) = K(t+\tau, s) - K(t, s)$$

We choose now a piecewise constant function v ,

$$v(t) = v_k, \quad t \in [\tau_k, \tau_{k+1}).$$

We represent

$$\begin{aligned} w_{k+1} &= w_k + \int_0^{\tau} K(\tau_{k+1}, \tau_k + s)v(\tau_k + s) \, ds + \int_0^{\tau_k} F(\tau_k, s)v(s) \, ds \\ &= w_k + A_k v_k + \sum_{j=0}^{k-1} F_{k,j} v_j. \end{aligned} \quad (5)$$

Here

$$A_k = \int_0^\tau K(\tau_{k+1}, \tau_k + s) \, ds, \quad F_{k,j} = \int_{\tau_j}^{\tau_{j+1}} F(\tau_k, s) \, ds.$$

We want a rule for choosing the constant value v_k at each time τ_k . As suggested by the finite dimensional case, We choose

$$v_k = \arg \min \left\{ \|w_k + A_k v - \xi_{k+1}\|^2 + \alpha \tau \|v\|^2 \right\}$$

i.e.

$$v_k = -[\alpha \tau I + A_k^* A_k]^{-1} A_k^* [w_k - \xi_{k+1}] \quad (6)$$

(we recall that ξ_k is the observation).

We shall prove that the piecewise constant function $v(t)$ so constructed approximates the unknown input u in the following two cases: the convolution case, under the assumptions described in sect. 7; the case that the kernel $K(t, s)$ is Lipschitz continuous in t , uniformly for $0 \leq s \leq t \leq T$. If the kernel is merely continuous, with $\det K(t, t) \neq 0$ (see subsection 6 for the precise statement) we must replace the piecewise constant function v with the piecewise continuous function v defined by

$$v(t) = -\frac{1}{\alpha} [w(t) - \xi_k], \quad t \in [\tau_k, \tau_{k+1}]. \quad (7)$$

With this definition, the candidate approximation v of u is constructed by

$$w(t) = -\frac{1}{\alpha} \int_0^t K(t, s) [w(s) - \xi(s)] \, ds, \quad v(t) = -\frac{1}{\alpha} [w(t) - \xi(t)],$$

where

$$\xi(t) = \xi_k \quad t \in [\tau_k, \tau_{k+1}]. \quad (8)$$

We prove that this input v indeed approximates the unknown input u if τ , α and h converge to zero while respecting suitable conditions.

6. The nonconvolution equation

We state first the assumption on the kernel K :

Assumption 1. The kernel is a square $n \times n$ matrix, continuous for $0 \leq s \leq t \leq T$ and satisfies

$$K(t, t) = I \quad t \in [0, T].$$

Moreover, $t \rightarrow K(t, s)$ is differentiable for a.e. $s \in [0, t]$ and $H(t, s) = K_t(t, s)$ satisfies

$$\|H(t, s)\| \leq L(s), \quad 0 \leq s \leq t \leq T, \quad \lim_{h \rightarrow 0} \int_0^t \|K_t(t+h, s) - K_t(t, s)\|^p \, ds = 0 \quad (9)$$

where $L(s) \in L^p(0, T)$, $p > 1$.

If these conditions hold then the solution u is unique and we can show the following result, where $\gamma = (p - 1)/p$ if $p < +\infty$, $\gamma = 1$ if $p = \infty$:

Theorem 3 *Let*

$$\tau, \alpha \text{ and } h \text{ converge to zero; } \lim \frac{\tau^\gamma}{\alpha} = 0, \quad \lim \frac{h}{\alpha} = 0. \quad (10)$$

Then: 1) *If u is measurable and bounded then the sequence of the functions v converges to u in $L^p(0, T)$ for every $p \in [1, +\infty)$.* 2) *If $u \in C(a, b)$ and $[a, b] \subseteq (0, T]$ then the convergence is uniform on $[a, b]$.* 3) *if $u \in C(0, T)$ then for every $\sigma > 0$ the convergence is uniform on $[\sigma, T]$.* 4) *If $u \in C(0, T)$ and if, furthermore, $u(0) = 0$ then the convergence is uniform on $[0, T]$.*

The first step in the proof of Theorem 3 is the proof that w tracks y . We sketch this part of the proof.

It is clear that, under **Assumption 1**, the output y is Hölder continuous: there exists a number $M_0 > 0$ and $\gamma \in (0, 1]$ such that for $t \in [\tau_k, \tau_{k+1})$ we have $\|y(t) - y(\tau_k)\| \leq M_0 \tau^\gamma$. Let us introduce the function $\phi(t) = \xi(t) - y(t)$, $t \in [\tau_k, \tau_{k+1})$, $\|\phi(t)\| \leq M_0(h + \tau^\gamma)$ (the function $\xi(t)$ is the one defined in (8)) so that

$$\|w(t) - \xi(t)\| \leq \|w(t) - y(t)\| + \|\phi(t)\| \leq \|w(t) - y(t)\| + C(\tau^\gamma + h).$$

Hence, v and w solve

$$v(t) = \frac{y(t) - w(t)}{\alpha} + \frac{\phi(t)}{\alpha}$$

$$y(t) - w(t) = -\frac{1}{\alpha} \int_0^t K(t, s)[y(s) - w(s)] \, ds - \frac{1}{\alpha} \int_0^t K(t, s)\phi(s) \, ds + y(t).$$

Let

$$e(t) = y(t) - w(t).$$

We prove firstly an estimate for $e(t)$ in terms of the parameters τ , h and α .

The function e is a.e. differentiable, with

$$e'(t) = -\frac{1}{\alpha}e(t) - \frac{1}{\alpha} \int_0^t H(t, s)e(s) \, ds - \frac{\phi(t)}{\alpha} - \frac{1}{\alpha} \int_0^t H(t, s)\phi(s) \, ds + y'(t).$$

It follows

$$\begin{aligned} e(t) &= \int_0^t e^{-(t-s)/\alpha} y'(s) \, ds - \int_0^t e^{-(t-s)/\alpha} \frac{\phi(s)}{\alpha} \, ds \\ &\quad - \int_0^t e^{-(t-s)/\alpha} \frac{1}{\alpha} \int_0^s H(s, r)e(r) \, dr \, ds \\ &\quad - \int_0^t e^{-(t-s)/\alpha} \frac{1}{\alpha} \int_0^s H(s, r)\phi(r) \, dr \, ds. \end{aligned}$$

We use now *boundedness* of u to obtain

$$\|e(t)\| \leq \mathcal{C}[\alpha + \tau^\gamma + h] + \int_0^t L(r) \|e(r)\| dr.$$

It follows:

$$0 \leq \|e(t)\| \leq z_\alpha(t) \leq \mathcal{M}[\alpha + \tau^\gamma + h]. \quad (11)$$

The constant \mathcal{M} *does not* depend on τ , α and h . this implies that e converges to zero *uniformly* on $[0, T]$: Once that this is known, the proof of Theorem 3 is in two steps: we first prove that v converges weakly to u and then we use the compactness properties of the Volterra operator so to prove norm convergence.

6.1. Explicit convergence estimate

It is not possible to give convergence estimates without “a priori” information on the unknown input u . In order to give convergence estimates, we assume that u is Hölder continuous,

$$\|u(t) - u(s)\| \leq \mathcal{C}|t - s|^\eta.$$

Furthermore we assume that K is of class C^1 on the triangle $0 \leq s \leq t \leq T$.

We have:

Theorem 4 *Let u be Hölder continuous of exponent η on $[0, T]$ and let K be Lipschitz continuous. There exists a number M such that for every $\sigma > 0$, τ , h , α the following convergence estimate holds:*

$$\|v(t) - u(t)\| < M_\sigma \left\{ \frac{\tau^\gamma + h}{\alpha} + \alpha^\eta + e^{-\sigma/\alpha} \right\}, \quad t \in [\sigma, T].$$

7. Convolution equations

We consider now the case that the kernel is a function of the difference of the arguments, $K(t, s) = K(t - s)$. In this case we assume that the kernel is *scalar* and of class $L^1(0, T)$. For simplicity of presentation, in order to have a continuous output, in this talk we assume u piecewise continuous. The general case of $u \in L^2(0, T)$ can be found in [3].

Experience with the previous cases suggests that now we can choose $v(t)$ as

$$v(t) = -\frac{w(\tau_k) - \xi_k}{\alpha} \quad t \in [\tau_k, \tau_{k+1}).$$

We study directly the error between v and u . Hence now the *error function* is $e(t) = v(t) - u(t)$. It solves

$$\alpha e = -\alpha u + [K * u - K * v]^\tau + \theta^\tau \quad (12)$$

where ${}^\tau$ denotes sampling,

$$f^\tau(t) = f(\tau_k) \quad t \in [\tau_k, \tau_{k+1}).$$

We want to compute the Laplace transform of both sides. We recall that we are working on a finite time interval. However, we can extend $K(t)$ to $[0, +\infty)$ so to have a Laplace transform and we can let $e(t)$ be defined by (12) for every $t > 0$. Let ν_C be the abscissa of convergence of the Laplace transform $\hat{K}(\lambda)$ of $K(t)$.

In order to compute the Laplace transform of both sides we need a formula for the transfer function from a *sample data input to the samples of a convolution*. Let $\hat{K}(\tau, \lambda)$ be such transfer function. It turns out that

$$\hat{K}(\tau, \lambda) = e^{-\lambda\tau} \sum_{n=0}^{+\infty} K_n e^{-\lambda\tau n}, \quad K_n = \int_{\tau n}^{\tau(n+1)} K(s) ds.$$

This is the fundamental object of our study.

We list now three conditions on the kernel K :

- (HP1) there exist *positive* numbers γ_1 , M_1 and $R > \nu_c$ such that $|\hat{K}(\lambda)| \leq \frac{M_1}{|\lambda|^{\gamma_1}}$ for $|\lambda| > R$;
- (HP2) there exist *positive* numbers γ_2 , M_2 and $R > \nu_c$ such that $|\hat{K}(\lambda)| \geq \frac{M_2}{|\lambda|^{\gamma_2}}$ for $|\lambda| > R$.
- (HP3) We assume that there is a sector $S_{r,\theta} = \{\lambda \in \mathbb{C}, |\lambda| < r, |\operatorname{Arg} \lambda| > \theta\}$ and a positive number $\nu_S \geq \nu_C$ such that $\Re \lambda > \nu_S \Rightarrow \hat{K}(\lambda) \notin S_{r,\theta}$.

Remark 5 We observe that the previous assumptions (HP1)–(HP3) are satisfied by a very large class of kernels, in particular Abel kernels $1/t^\gamma$, $[0 \leq \gamma < 1]$ and piecewise regular kernels, as proved in [3]. The sector condition (HP3) has been formulated with a condition $\liminf_{t \rightarrow 0^+} K(t) \geq 0$ in mind. In concrete applications it may require to work with $-K$ instead then with K . ■

Conditions (HP1) and (HP2) justify the use of a formula originally due to Poisson, from which we obtain:

$$\hat{K}(\tau, \lambda) = \sum_{n=-\infty}^{n=+\infty} \left[\frac{1 - e^{-\lambda\tau}}{\lambda\tau + 2n\pi i} \right] \hat{K}\left(\lambda + \frac{2n\pi i}{\tau}\right). \quad (13)$$

Formula (13) shows that $\hat{K}(\tau, \lambda)$ is periodic of period $2\pi i/\tau$ and moreover, on each horizontal strip $[2k\pi i/\tau, 2(k+1)\pi i/\tau]$ one of the term is dominant. In particular, in the strip $[-\pi i/\tau, \pi i/\tau]$ the dominant term is the one of index 0.

Combining the three assumptions on K we have:

Theorem 6 *Assume conditions **(HP0)**, **(HP1)**, and **(HP3)**. Then, there exist numbers $\tau_0 > 0$, $L > 0$ and a sector $S_{\tilde{r}, \tilde{\theta}}$ such that if $\tau \in (0, \tau_0)$ and if λ is such that $\Re \lambda > \nu_S$, then $\hat{K}(\tau, \lambda) \in S_{\tilde{r}, \tilde{\theta}} \implies \Re \hat{K}(\tau, \lambda) > -L\tau^{\gamma_1}$.*

This is the crucial result needed in the proof of the following consistency result:

Theorem 7 *Assume conditions **(HP0)**–**(HP3)**. Then, for every $\epsilon > 0$, there exist α_0 , τ_0 , h_0 such that*

$$\|e_{\alpha_0, \tau, \xi}\|_2 < \epsilon, \quad \forall \tau < \tau_0, \quad \forall \xi : \|\xi\|_2 < h_0.$$

As usual, convergence estimates can be obtained provided that we have “a priori” informations on the regularity of the unknown input u . For example,

Theorem 8 *Assume conditions **(HP0)**–**(HP3)** and let $u \in W^{1,2}(0, T)$. Then, there exist $M \geq 0$, δ_1 , $\delta_2 > 0$ such that*

$$\|e_{\alpha, \tau, \xi}\|_2 \leq M \left[\alpha^{\frac{1}{1+2\gamma_2}} + \frac{\sqrt{\tau}}{\alpha} + \frac{h}{\alpha} \right]$$

provided that $\tau, \alpha, \xi : \tau^{\gamma_1}/\alpha < \delta_1, \alpha < \delta_2, \|\xi\| < h$.

Finally we note that in the proof of the previous results the fact that we are working on a finite time interval is explicitly used, in spite of the frequency domain nature of the method. The analysis of problems on $[0, +\infty)$ is still under study. We state a preliminary result which holds when condition **(HP2)** holds in the following more restrictive form:
(HP2+) There exist positive numbers γ_2 and \tilde{M}_2 such that

$$|\hat{K}(\lambda)| \geq \frac{\tilde{M}_2}{1 + |\lambda|^{\gamma_2}}.$$

for any λ such that $\Re \lambda > 0$. We have:

Theorem 9 *Let $K \in L^1(0, +\infty)$ and assume conditions **(HP1)**, **(HP2+)**, and **(HP3)**. Assume moreover, $\nu_S = 0$. Then, for every $\epsilon > 0$, there exist α_0 , τ_0 , h_0 such that*

$$\|e_{\alpha_0, \tau, \xi}\|_{L^2(0, +\infty)} < \epsilon, \quad \forall \tau < \tau_0, \quad \forall \xi : \|\xi\|_2 < h_0.$$

References

- [1] F. Fagnani, V. Maksimov, L. Pandolfi, A recursive deconvolution approach to disturbance reduction, Rapporto interno n. 15, Dip. di Matematica, Pol. Torino, 2001.
- [2] F. Fagnani, L. Pandolfi, A singular perturbation approach to a recursive deconvolution problem, *SIAM J. Control Optim.* **40** 1384–1405, 2002.
- [3] F. Fagnani, L. Pandolfi, A recursive algorithm for the approximate solution of Volterra integral equations of first kind, Rapporto interno n. 43, Dip. di Matematica, Pol. Torino, 2001.
- [4] A. Favini, V. Maksimov, L. Pandolfi, A deconvolution problem related to a singular system, submitted
- [5] R. Gorenflo, S. Vessella, *Abel Integral Equations*, L.N. Mathematics 1461, Springer-Verlag, Berlin, 1991.
- [6] S. Komada, M. Ishida, K. Ohnishi, T. Hori, Disturbance observer-based motion control of direct drive motors, *IEEE transaction on energy conversion*, **6** (1991) 553–559.
- [7] N.N. Krasovskii, A.I. Subbotin, *Game-Theoretical Control Problems*, Springer Verlag, New York – Berlin, 1988.
- [8] A.V. Kryazhimskii, Yu.S. Osipov, *Inverse Problems for Ordinary Differential Equations: Dynamical Solutions*, Gordon and Breach, London, 1995.
- [9] V. Maksimov, L. Pandolfi, Dynamical reconstruction of inputs for contraction semigroup systems: the boundary input case, *J. Optim. Theory Appl.*, **103** 401–420, 1999.
- [10] V. Maksimov, L. Pandolfi, The problem of dynamical reconstruction of Dirichlet boundary control in semilinear hyperbolic systems, *J. Inverse ill-posed problems*, **8** 1–22, 2000.
- [11] V. Maksimov, L. Pandolfi, On a dynamical identification of controls in nonlinear time-lag systems, *IMA J. Math. Control Inf.* **19** 173–184, 2002.
- [12] A.S. Morse, Structural invariants of linear multivariable systems, *SIAM J. Control*, **11** (1973), pp. 446–465.
- [13] Yu. Osipov, L. Pandolfi, V. Maksimov, Problems of dynamical reconstruction and robust boundary control: the case of Dirichlet boundary conditions, *J. Inv. Ill-posed Problems*, **9** 149–162, 2001.
- [14] L. Pandolfi, Disturbance decoupling and invariant subspaces for delay systems, *Appl. Math. Optim.*, **14** 55–72, 1986.
- [15] E. Vasil'eva, V. Maksimov, On the dynamical reconstruction of control in differential equations with memory, *Diff. Equat.* **35** 815–824, 1999.

DETERMINING A SEMILINEAR PARABOLIC PDE FROM FINAL DATA *

Luis A. Fernández

Dep. de Matemáticas, Estadística y Computación

Universidad de Cantabria. 39071 – Santander (SPAIN)

lafernandez@unican.es

Cecilia Pola

Dep. de Matemáticas, Estadística y Computación

Universidad de Cantabria. 39071 – Santander (SPAIN)

polac@unican.es

Abstract Given the following boundary value problem

$$\begin{cases} y_t - \Delta y + a(x)b(y) = u\chi_{\omega_{in}} & \text{in } \Omega \times (0, T), \\ y = 0 & \text{on } \partial\Omega \times (0, T), \\ y(0) = y_0 & \text{in } \Omega, \end{cases}$$

we are concerned with the following *inverse problem*: assuming that for each source term $u(x, t) \in L^2(\omega_{in} \times (0, T))$ the solution at the final time $y(x, T)$ is known over ω_{out} , we want to determine the nonlinear term $a(x)b(y)$ over $\Omega \times \mathbb{R}$. We suppose that ω_{in} and ω_{out} are (probably small) fixed open subsets of Ω with $\omega_{in} \cap \omega_{out} \neq \emptyset$ and $\chi_{\omega_{in}}$ denotes the characteristic function of ω_{in} .

Keywords: inverse problem, semilinear parabolic PDE, parameter estimation.

1. Introduction

Let us consider the following boundary value problem, associated with a semilinear parabolic equation

$$\begin{cases} y_t(x, t) - \Delta y(x, t) + a(x)b(y(x, t)) = u(x, t)\chi_{\omega_{in}}(x) & \text{in } Q, \\ y(x, t) = 0 & \text{on } \Sigma, \\ y(x, 0) = y_0(x) & \text{in } \Omega, \end{cases} \quad (1)$$

*This work was partially supported by DGES (Spain). Project PB 98 – 0193.

where Ω is a bounded open subset of \mathbb{R}^n with a boundary $\partial\Omega$ of class C^2 , $Q = \Omega \times (0, T)$, $\Sigma = \partial\Omega \times (0, T)$ and $T > 0$.

Given the functions a , b , u and y_0 , the classical direct problem consists of determining $y(x, t)$ over Q . Here we are concerned with the following *inverse problem*: assuming that the initial datum y_0 is known and that for each source term $u \in L^2(\omega_{in} \times (0, T))$ the solution at the final time $y(x, T)$ is known over $\omega_{out} \subset \Omega$, we want to determine the nonlinear term $a(x)b(y)$.

Our work is somehow related to the recovery of a nonlinear term $f(x, t, y)$ through the knowledge of the Dirichlet-to-Neumann map associated to the semilinear PDE, see [4, Chapter 9] and the bibliography therein.

On the other hand, in [2], we have studied the case where the source terms u are acting on the whole domain Q (i.e. $\omega_{in} = \Omega$). In fact, in that case, we have proved the simultaneous identification of the nonlinearities A , \bar{b} and c appearing in the quasilinear parabolic equation

$$y_t - \operatorname{div}(A(y)\nabla y + \bar{b}(y)) + c(y) = u(x, t) \quad \text{in } Q,$$

that are unique in an appropriate class.

The main objective of this work is to deal with the case $\omega_{in} \neq \Omega$. To this end, we restrict ourselves to the semilinear PDE given in (1), although the linear part of the operator can be taken in a more general way. In the first part of the work, we prove the uniqueness of $a(x)b(y)$ assuming that a is analytic on $\overline{\Omega}$ without changing the sign and $b \in C^1(\mathbb{R})$ with $b' \in L^\infty(\mathbb{R})$. This property can be deduced by using the approximate controllability for this type of equations, when the function u is viewed as a control.

In the second part, we present a finite-dimensional optimization problem that allows the practical reconstruction of $a(x)b(y)$. In order to illustrate the applicability of the method, some numerical results are shown in the one-dimensional space case (i.e. $n = 1$). The influence of noise in the data will be also taken into account.

2. Uniqueness of the nonlinear term

Let us recall the space

$$W(0, T) = \{y \in L^2(0, T; H_0^1(\Omega)) : y_t \in L^2(0, T; H^{-1}(\Omega))\},$$

where $H_0^1(\Omega)$ denotes the classical Sobolev space and $H^{-1}(\Omega)$ its dual.

Let us introduce the hypotheses that will be assumed on the nonlinear term of the semilinear parabolic operator:

- H1) a is a real analytic function on $\overline{\Omega}$ and $a(x) \geq 0 \quad \forall x \in \overline{\Omega}$.
- H2) $b \in C^1(\mathbb{R})$ with $b' \in L^\infty(\mathbb{R})$.

The sign condition in H1) is just made to fix ideas: the important point here is that a does not change the sign in Ω (see the end of the proof of Theorem 2).

For the sake of brevity, we designate $\mathcal{C} = \{(a, b) \text{ verifying } H1) - H2)\}$.

Given $(a, b) \in \mathcal{C}$, $u \in L^2(\omega_{in} \times (0, T))$ and $y_0 \in L^2(\Omega)$, it is well known that the problem (1) has a unique solution y in $W(0, T)$, see [5, Theorem 6.7, pp. 466]. Taking into account that the initial datum y_0 will remain fixed through the rest of the paper, we will designate by $y_{u,a,b}$ the unique solution of problem (1) in $W(0, T)$.

Let us now recall an approximate controllability result for parabolic equations:

Theorem 1 *Let us fix $y_0 \in L^2(\Omega)$, $(a, b) \in \mathcal{C}$ and ω_{in} a nonempty open subset of Ω . Then, the set $R(T) = \{y_{u,a,b}(x, T) : u \in L^2(\omega_{in} \times (0, T))\}$ is dense in $L^2(\Omega)$.*

This is a particular case of a more general result (see for instance [1] and [3]). At this level, the assumption H1) on a can be weakened to $a \in L^\infty(\Omega)$.

With the help of this controllability result, we can establish the following uniqueness theorem:

Theorem 2 (Uniqueness) *Let us fix $y_0 \in L^2(\Omega)$ and $\omega_{in}, \omega_{out}$ two open subsets of Ω such that $\omega_{in} \cap \omega_{out} \neq \emptyset$. Suppose that there exist two pairs $(a, b), (\tilde{a}, \tilde{b}) \in \mathcal{C}$ such that $y_{u,a,b}(T) = y_{u,\tilde{a},\tilde{b}}(T)$ in ω_{out} for each $u \in L^2(\omega_{in} \times (0, T))$. Then, $a(x)b(s) = \tilde{a}(x)\tilde{b}(s)$ for all $(x, s) \in \Omega \times \mathbb{R}$.*

Proof. Given $u, v \in L^2(\omega_{in} \times (0, T))$ and $\lambda \in (0, 1)$, let us consider $y_{u+\lambda v, a, b}$ and $y_{u, a, b}$. Moreover we define $z_{\lambda, a, b} = (y_{u+\lambda v, a, b} - y_{u, a, b})/\lambda$. Analogously, from $y_{u+\lambda v, \tilde{a}, \tilde{b}}$ and $y_{u, \tilde{a}, \tilde{b}}$ we introduce $z_{\lambda, \tilde{a}, \tilde{b}} = (y_{u+\lambda v, \tilde{a}, \tilde{b}} - y_{u, \tilde{a}, \tilde{b}})/\lambda$. By hypothesis we know that $y_{u, a, b}(T) = y_{u, \tilde{a}, \tilde{b}}(T)$ in ω_{out} and $y_{u+\lambda v, a, b}(T) = y_{u+\lambda v, \tilde{a}, \tilde{b}}(T)$ in ω_{out} for each λ . Consequently, $z_{\lambda, a, b}(T) = z_{\lambda, \tilde{a}, \tilde{b}}(T)$ in ω_{out} . By using the Mean Value Theorem, it can be proved that $z_{\lambda, a, b} \rightarrow z_{v, a, b}$ in $W(0, T)$ as $\lambda \rightarrow 0$, where $z_{v, a, b}$ is the unique solution in $W(0, T)$ of the problem

$$\begin{cases} z_t(x, t) - \Delta z(x, t) + a(x)b'(y_{u, a, b})(x)z(x, t) = v(x, t)\chi_{\omega_{in}}(x) & \text{in } Q, \\ z(x, t) = 0 & \text{on } \Sigma, \\ z(x, 0) = 0 & \text{in } \Omega. \end{cases}$$

Thanks to H2), let us point out that $a(x)b'(y_{u, a, b}(x, t)) \in L^\infty(Q)$. Hence, it is clear that previous problem is well posed in $W(0, T)$.

The same argumentation leads us to show that $z_{\lambda,\tilde{a},\tilde{b}} \rightarrow z_{v,\tilde{a},\tilde{b}}$ in $W(0,T)$ as $\lambda \rightarrow 0$, where $z_{v,\tilde{a},\tilde{b}}$ is the unique solution in $W(0,T)$ of the problem

$$\begin{cases} z_t(x,t) - \Delta z(x,t) + \tilde{a}(x)\tilde{b}'(y_{u,\tilde{a},\tilde{b}})z(x,t) = v(x,t)\chi_{\omega_{in}}(x) & \text{in } Q, \\ z(x,t) = 0 & \text{on } \Sigma, \\ z(x,0) = 0 & \text{in } \Omega. \end{cases}$$

Furthermore, $z_{v,a,b}(T) = z_{v,\tilde{a},\tilde{b}}(T)$ in ω_{out} . Now, we introduce $w = z_{v,a,b} - z_{v,\tilde{a},\tilde{b}}$ that satisfies $w(T) = 0$ in ω_{out} and it can be viewed as the unique solution in $W(0,T)$ of the linear problem

$$\begin{cases} w_t - \Delta w + ab'(y_{u,a,b})w = (\tilde{a}\tilde{b}'(y_{u,\tilde{a},\tilde{b}}) - ab'(y_{u,a,b}))z_{v,\tilde{a},\tilde{b}} & \text{in } Q, \\ w(x,t) = 0 & \text{on } \Sigma, \\ w(x,0) = 0 & \text{in } \Omega. \end{cases} \quad (2)$$

For each $\phi \in L^2(\omega_{out})$, we consider the following adjoint problem:

$$\begin{cases} -p_t(x,t) - \Delta p(x,t) + a(x)b'(y_{u,a,b}(x,t))p(x,t) = 0 & \text{in } Q, \\ p(x,t) = 0 & \text{on } \Sigma, \\ p(x,T) = \phi(x)\chi_{\omega_{out}}(x) & \text{in } \Omega, \end{cases} \quad (3)$$

whose unique solution p_ϕ also belongs to $W(0,T)$.

Multiplying the equation (2) by p_ϕ , integrating by parts and taking into account that $w(T)p(T) = 0$ in Ω , we obtain

$$\int_Q (\tilde{a}(x)\tilde{b}'(y_{u,\tilde{a},\tilde{b}}(x,t)) - a(x)b'(y_{u,a,b}(x,t)))z_{v,\tilde{a},\tilde{b}}(x,t)p_\phi(x,t)dxdt = 0, \quad (4)$$

for all $v \in L^2(\omega_{in} \times (0,T))$ and all $\phi \in L^2(\omega_{out})$.

Finally, let us introduce a new boundary problem associated with the previous ones, given by

$$\begin{cases} -q_t - \Delta q + \tilde{a}\tilde{b}'(y_{u,\tilde{a},\tilde{b}})q = (\tilde{a}\tilde{b}'(y_{u,\tilde{a},\tilde{b}}) - ab'(y_{u,a,b}))p_\phi & \text{in } Q, \\ q(x,t) = 0 & \text{on } \Sigma, \\ q(x,T) = 0 & \text{in } \Omega. \end{cases} \quad (5)$$

Once more, multiplying the PDE in (5) by $z_{v,\tilde{a},\tilde{b}}$, integrating by parts, taking into account (4) and the PDE satisfied by $z_{v,\tilde{a},\tilde{b}}$, we derive

$$\int_Q v(x,t)\chi_{\omega_{in}}(x)q(x,t)dxdt = 0, \text{ for all } v \in L^2(\omega_{in} \times (0,T)). \quad (6)$$

Therefore, $q(x, t) = 0$ in $\omega_{in} \times (0, T)$ and using the PDE in (5) we get

$$(\tilde{a}(x)\tilde{b}'(y_{u,\tilde{a},\tilde{b}}(x, t)) - a(x)b'(y_{u,a,b}(x, t)))p_\phi(x, t) = 0, \quad (7)$$

in $\omega_{in} \times (0, T)$, for all $\phi \in L^2(\omega_{out})$.

In particular, taking $t = T$ in (7), we have

$$(\tilde{a}(x)\tilde{b}'(y_{u,\tilde{a},\tilde{b}}(x, T)) - a(x)b'(y_{u,a,b}(x, T)))\phi(x)\chi_{\omega_{out}}(x) = 0 \quad \text{in } \omega_{in},$$

for all $\phi \in L^2(\omega_{out})$. Therefore,

$$\tilde{a}(x)\tilde{b}'(y_{u,\tilde{a},\tilde{b}}(x, T)) = a(x)b'(y_{u,a,b}(x, T)) \quad \text{in } \omega_{in} \cap \omega_{out}. \quad (8)$$

Taking into account that $y_{u,a,b}(T) = y_{u,\tilde{a},\tilde{b}}(T)$ in ω_{out} for each u in $L^2(\omega_{in} \times (0, T))$ and that Theorem 1 implies the density of

$$\{y_{u,a,b}(T)|_{\omega_{in} \cap \omega_{out}} : u \in L^2(\omega_{in} \times (0, T))\}$$

in $L^2(\omega_{in} \cap \omega_{out})$, for each $s \in \mathbb{R}$ we can select a sequence $\{u_m\} \subset L^2(\omega_{in} \times (0, T))$ such that as m tends to $+\infty$

$$y_{u_m,a,b}(T) \longrightarrow s \quad \text{in } L^2(\omega_{in} \cap \omega_{out}).$$

Using the identity (8) for each u_m and passing to the limit in m , we arrive to $\tilde{a}(x)\tilde{b}'(s) = a(x)b'(s)$ for all $(x, s) \in (\omega_{in} \cap \omega_{out}) \times \mathbb{R}$. When $\tilde{a}(x) \neq 0$ and $b'(s) \neq 0$, we derive the existence of a real constant C_1 such that

$$\frac{\tilde{a}(x)}{\tilde{a}(s)} = \frac{\tilde{b}'(s)}{b'(s)} = C_1 \quad \forall x \in \omega_{in} \cap \omega_{out}, \forall s \in \mathbb{R}.$$

Due to the unique analytic continuation property, $a(x) = C_1\tilde{a}(x)$ for all $x \in \Omega$. Moreover, there exists a real constant $C_2 \in \mathbb{R}$ such that $\tilde{b}(s) = C_1b(s) + C_2$ for all $s \in \mathbb{R}$. Let us point out that $\tilde{a}(x)\tilde{b}(s) = a(x)b(s) + C_2\tilde{a}(x)$ for all $(x, s) \in \Omega \times (0, T)$.

To show that $C_2 = 0$, we fix $u \in L^2(\omega_{in} \times (0, T))$ and introduce $\zeta = y_{u,a,b} - y_{u,\tilde{a},\tilde{b}}$. Arguing as at the beginning of the proof, we derive that ζ is the unique solution of the following linear problem:

$$\begin{cases} \zeta_t(x, t) - \Delta\zeta(x, t) + a(x)c(x, t)\zeta(x, t) = C_2\tilde{a}(x) & \text{in } Q, \\ \zeta(x, t) = 0 & \text{on } \Sigma, \\ \zeta(x, 0) = 0 & \text{in } \Omega, \end{cases} \quad (9)$$

where $c(x, t) = \int_0^1 b'(sy_{u,a,b}(x, t) + (1-s)y_{u,\tilde{a},\tilde{b}}(x, t))ds$. Moreover, by hypothesis we know that $\zeta(x, T) = 0$ in ω_{out} . Once more, given $\phi \in L^2(\omega_{out})$, we consider \hat{p} the unique solution of the adjoint problem

$$\begin{cases} -\hat{p}_t(x, t) - \Delta \hat{p}(x, t) + a(x)c(x, t)\hat{p}(x, t) = 0 & \text{in } Q, \\ \hat{p}(x, t) = 0 & \text{on } \Sigma, \\ \hat{p}(x, T) = \phi(x)\chi_{\omega_{out}}(x) & \text{in } \Omega. \end{cases} \quad (10)$$

Multiplying by \hat{p} the PDE satisfied by ζ , integrating by parts and using that $\hat{p}(T)\zeta(T) = 0$ in Ω , we deduce

$$C_2 \int_Q \tilde{a}(x)\hat{p}(x, t)dxdt = 0. \quad (11)$$

Choosing $\phi(x) \geq 0$ for all $x \in \omega_{out}$, $\phi \not\equiv 0$, the Maximum Principle implies $\hat{p}(x, t) \geq 0$ in Q , $\hat{p} \not\equiv 0$. Noticing that $\tilde{a}(x) \geq 0$ in Ω , the equality (11) implies $C_2 = 0$. ■

Remarks.

- i) It follows from the proof that the functions $a(x)$ and $b(y)$ can be determined separately in a unique way, up to a multiplicative constant, because $a(x)b(y) = \frac{a(x)}{C_1}C_1b(y)$ for all $C_1 > 0$.
- ii) Let us point out that the conclusion of Theorem 2 remains valid if it is supposed that the relation $y_{u,a,b}(T) = y_{u,\tilde{a},\tilde{b}}(T)$ in ω_{out} only holds when $u \in L^2(\omega_{in} \times [T - \delta, T])$, where $\delta \in (0, T)$ is fixed. This is a consequence of the fact that the controllability result also holds in this case, i.e. under the conditions of Theorem 1, the set $R(T) = \{y_{u,a,b}(x, T) : u \in L^2(\omega_{in} \times [T - \delta, T])\}$ is dense in $L^2(\Omega)$, (see [3, Remark 3.1 (c)]). Here, $y_{u,a,b}$ denotes the unique solution of problem (1), when the right hand term of the PDE is given by $u(x, t)\chi_{\omega_{in} \times [T - \delta, T]}(x, t)$.
- iii) Exactly the same technique can be used to prove the uniqueness of the term $a(x)y + b(y)$ for the Dirichlet boundary problem associated to the PDE $y_t - \Delta y + a(x)y + b(y) = u\chi_{\omega_{in}}$, under the assumptions that a is analytic on $\bar{\Omega}$ and H2) (no sign condition has to be imposed on a).

3. The identification process. Numerical experiments.

In this section, problem (1) is considered in the one-dimensional space case (i.e. $n = 1$). We are concerned with getting a numerical approximation $\check{a}(x)\check{b}(y)$ of the nonlinear term $a(x)b(y)$ from a finite number of observations η_{ij} . Let us consider (1) for a finite set of source terms $\{u_j\}_{j=1}^{n_u}$. For each u_j , we have measurements η_{ij} of $y_{u_j,a,b}(x_i, T)$ at some points $x_i \in \omega_{out}$.

For our experiments we have chosen $y_0(x) = 0.5x(4 - x)$, $T = 1$ and $\Omega = (0, 4)$ with the grid $\{x_i = i/10\}_{i=0}^{40}$. The data η_{ij} correspond with the nodes x_i in $\omega_{out} = (0.4 - 1.e - 12, 2 + 1.e - 12)$ and they were obtained solving (1) with six of the following source terms:

$$u_j(x, t) = \begin{cases} -50(i_j - 2k_j) \sin(i_j \pi t) \sin(2k_j \pi x) & \text{if } x \in (0, 0.5), \\ 0 & \text{if } x \in [0.5, 4], \end{cases} \quad (12)$$

for $i_j, k_j \in \{1, 2, \dots, 15\}$. Note that $\omega_{in} = (0, 0.5)$ and only the node $x_i = 0.4$ belongs to $\omega_{in} \cap \omega_{out}$.

The first step of the identification process is related with the choice of the interval (for the variable y) where we are going to recover the non-linearity $a(x)b(y)$. Our working interval is an enlargement of the interval defined by the data, $I = [\eta_{min} = \min \eta_{ij}, \eta_{max} = \max \eta_{ij}]$, by taking a safety barrier, $M > 0$, in both extremes: $I_w = [\eta_{min} - M, \eta_{max} + M]$. In our calculations we took an equidistant grid $\{y_i\}_{i=0}^{n_y}$ in I_w with a mesh size $h = (\eta_{max} - \eta_{min})/10$.

The reconstruction of $a(x)b(y)$ consists in determining two vectors $\check{a} = (\check{a}_0, \dots, \check{a}_{n_x}) \in \mathbb{R}^{n_x+1}$ and $\check{b} = (\hat{b}_0, \dots, \hat{b}_{n_y}) \in \mathbb{R}^{n_y+1}$, where \check{a}_i and \hat{b}_i are the coefficients of the approximations

$$\check{a}(x) = \sum_{i=0}^{n_x} \check{a}_i B_i(x) \quad \text{and} \quad \check{b}(y) = \sum_{i=0}^{n_y} \hat{b}_i \hat{B}_i(y), \quad (13)$$

being B_i and \hat{B}_i piecewise linear B-splines verifying

$$B_i(x_{2j}) = \delta_{ij} \text{ for } i, j = 0, \dots, n_x \quad \text{and} \quad \hat{B}_i(y_j) = \delta_{ij} \text{ for } i, j = 0, \dots, n_y.$$

Let us point out that we use double mesh size for the function \check{a} than for the discretization of Ω . For inverse problems, some experiments indicate that it is better to take a bigger mesh size for the spatially varying parameters than for the solution y .

Now we are ready to state the finite-dimensional problem for recovering $a(x)b(y)$. Following an output least squares method with regularization, we consider the minimization problem

$$\min_{(\hat{a}, \hat{b}) \in U_{ad}} J(\hat{a}, \hat{b}), \quad (14)$$

where

$$\begin{aligned} J(\hat{a}, \hat{b}) = & \frac{1}{2} \sum_{j=1}^{n_u} \int_{\omega_{out}} (y_{u_j, \check{a}, \check{b}}(x, T) - \eta_j(x))^2 dx + \\ & + \gamma \left(\int_{\Omega \times I_w} \|\nabla(\check{a}(x)\check{b}(y))\|_2^2 dx dy + \epsilon \left(\int_{\Omega} (\check{a}(x))^2 dx + \int_{I_w} (\check{b}(y))^2 dy \right) \right), \end{aligned} \quad (15)$$

being $\eta_j(x)$ a continuous observation which interpolates the data η_{ij} , $i = 4, \dots, 20$, $\gamma > 0$, $\epsilon \geq 0$ and U_{ad} the set of feasible vectors given by

$$U_{ad} = \{(\hat{a}, \hat{b}) \in \mathbb{R}^{n_x+1} \times \mathbb{R}^{n_y+1} : \hat{a}_i \geq 0, i = 0, \dots, n_x\}. \quad (16)$$

Hence, the objective function is the sum of a least squares term and a regularization term. If $\epsilon > 0$, the last term is useful for proving the coercivity of the functional. On the other hand, if $\epsilon = 0$, we also can obtain the coercivity of J by imposing some constraints, for instance: $\check{a}(x_0)^2 + \check{a}(x_{n_x})^2 \leq c_1$, $\check{b}(y_0)^2 + \check{b}(y_{n_y})^2 \leq c_2$, $\|\hat{a}\|^2 \geq c_3 > 0$ and $\|\hat{b}\|^2 \geq c_3 > 0$; but this kind of formulation needs some a priori knowledge about the functions to recover. In the simpler case $a(x) \equiv 1$, taking $\check{b}(y_0) = 0$, $\check{b}(y_{n_y}) = 0$, J is coercive on $U_{ad} = \mathbb{R}^{n_y+1}$ (see [2]).

The following result states the existence of minimizers.

Theorem 3 *There exists at least one solution to the problem (14).*

Proof. Because of J is continuous (see Appendix of [2]), coercive and the feasible set U_{ad} is non-empty and closed, the result follows. ■

Problem (14) was solved by using the subroutine E04UCF from NAG Library, that implements a sequential quadratic programming (SQP) algorithm. Each evaluation of the objective function was computed by taking (for each j) a linear spline which interpolates the values $\eta_{ij} - y_{u_j, \check{a}, \check{b}}(x_i, T)$, where $y_{u_j, \check{a}, \check{b}}(x_i, T)$ were obtained by solving (1) with a linearized Crank-Nicholson-Galerkin method; to this end, we took a semidiscrete approach with 39 piecewise linear finite elements for the discretization of the spatial domain and the nodes $\{iT/10\}_{i=0}^{10}$ for the discretization of the time variable.

About the stopping test, let us mention that the optimization algorithm terminates successfully if the following conditions are satisfied,

$$\frac{\|(\hat{a}^k - \hat{a}^{k-1}, \hat{b}^k - \hat{b}^{k-1})\|}{1 + \|(\hat{a}^k, \hat{b}^k)\|} < \sqrt{\epsilon}, \quad (17)$$

$$\frac{\|(\nabla J(\hat{a}^k, \hat{b}^k))_{FR}\|}{1 + \max\{1 + |J(\hat{a}^k, \hat{b}^k)|, \|(\nabla J(\hat{a}^k, \hat{b}^k))_{FR}\|\}} \leq \sqrt{\epsilon}, \quad (18)$$

$$(\nabla J(\hat{a}^k, \hat{b}^k))_L \geq 0,$$

where $(\nabla J(\hat{a}^k, \hat{b}^k))_{FR}$ is the vector with the components of $\nabla J(\hat{a}^k, \hat{b}^k)$ corresponding to the free variables (i.e. not fixed at the bounds) and $(\nabla J(\hat{a}^k, \hat{b}^k))_L$ is formed by the rest of the components of $\nabla J(\hat{a}^k, \hat{b}^k)$.

Concerning the bound conditions described in H1) and H2), let us point out that the lower bounds in (16) ensure $\check{a}^k(x) \geq 0$ for all k

and all $x \in \Omega$. On the other hand, from $J(\hat{a}^k, \hat{b}^k) \leq J(\hat{a}^0, \hat{b}^0)$, when $\int_{\Omega} (\check{a}^k(x))^2 dx \geq c_1 > 0$, it follows that

$$((\check{b}^k)'(y))^2 = \left(\sum_{i=0}^{n_y} \hat{b}_i \hat{B}'_i(y) \right)^2 \leq \frac{J(\hat{a}^0, \hat{b}^0)}{\gamma h c_1},$$

for all $y \notin \{y_0, \dots, y_{n_y}\}$ and for all k .

Our numerical experiments were carried out in MATLAB 6 on a DELL PowerEdge 2500 with 1 GB of memory and running under Red Hat Linux, taking $(0.5, \dots, 0.5)$ as the starting point.

To generate the observations η_{ij} , we solve (1) and take

$$\eta_{ij} = (1 + \delta_{ij}) y_{u_j, a, b}(x_i, T),$$

where δ_{ij} are uniformly distributed random numbers in $[-\delta, \delta]$, with δ denoting the noise level.

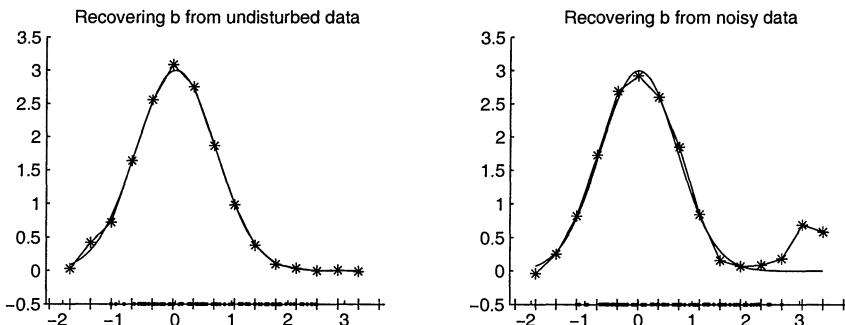
We have considered two examples.

Example 1 Knowing $a(x) \equiv 1$ in Ω , we try to recover $b(y) = 3\exp(-y^2)$ from the data obtained with six source terms u_j given by (12) with (i_j, k_j) taking the values $(1, 1), (1, 7), (2, 2), (2, 3), (2, 4)$ and $(2, 6)$. In Table 1 and Figure 1 we present the computed results obtained taking $\epsilon = 0$, $b(y_0) = 0$ and $b(y_{n_y}) = 0$ in the formulation of problem (14), $\epsilon = 1.e - 9$ for the stopping test. Table 1 summarizes numerical results corresponding to $\delta = 0$ (without explicit data perturbation) and $\delta = 0.1$ (noisy data). For each value of δ , the nodes in I_w are listed in the first column (4 nodes in the safety barrier and 11 nodes on the interval I). For each node y_i , we report the value of the numerical solution, $\check{b}(y_i)$, and the value of the exact solution, $b(y_i)$, in the second and third column respectively, and finally the relative error $|\check{b}(y_i) - b(y_i)| / \max(1, |b(y_i)|)$. At the bottom of the table we show the current values of the penalty parameter γ and the number of iterations required by the optimization routine.

Figure 1 contains two graphics. The nonlinear term $b(y)$ obtained from data with $\delta = 0$ and $\delta = 0.1$ appears on the left and on the right, respectively. Each graphic shows the exact solution b and the numerically identified solution \check{b} . We asterisk the values $\check{b}(y_i)$. Moreover, in each graphic we show the data η_{ij} on the axis. Remark that these observations depend on the source terms u_j . Here, the main point is to choose u_j in such a way that the corresponding observations fill as much as possible the interval where we want to recover the coefficient. Hence, it is our experience that, with a good choice, a small number of functions can be enough to recover the nonlinearity.

Table 1. Example 1

Undisturbed data ($\delta = 0$)				Noisy data ($\delta = 0.1$)			
y_i	$\dot{b}(y_i)$	$b(y_i)$	Error	y_i	$\dot{b}(y_i)$	$b(y_i)$	Error
-1.8744	0.0309	0.0894	5.854e-2	-1.9181	-0.0374	0.0757	1.132e-1
-1.5088	0.4179	0.3080	1.099e-1	-1.5362	0.2514	0.2832	3.181e-2
-1.1431	0.7212	0.8121	9.095e-2	-1.1544	0.8193	0.7914	2.788e-2
-0.7775	1.6366	1.6390	1.465e-3	-0.7725	1.7270	1.6518	4.556e-2
-0.4119	2.5548	2.5319	9.043e-3	-0.3906	2.6902	2.5755	4.455e-2
-0.0463	3.0884	2.9936	3.166e-2	-0.0088	2.9227	2.9998	2.568e-2
0.3194	2.7510	2.7091	1.546e-2	0.3731	2.6009	2.6101	3.526e-3
0.6850	1.8653	1.8765	5.932e-3	0.7550	1.8487	1.6965	8.968e-2
1.0506	0.9815	0.9948	1.336e-2	1.1369	0.8444	0.8238	2.064e-2
1.4163	0.3796	0.4037	2.407e-2	1.5188	0.1584	0.2988	1.404e-1
1.7819	0.0994	0.1254	2.599e-2	1.9006	0.0732	0.0810	7.731e-3
2.1475	0.0363	0.0298	6.480e-3	2.2825	0.0862	0.0164	6.976e-2
2.5131	-0.0011	0.0054	6.491e-3	2.6644	0.1814	0.0025	1.789e-1
2.8788	0.0102	0.0008	9.404e-3	3.0463	0.6878	0.0003	6.875e-1
3.2444	-0.0106	0.0001	1.072e-2	3.4281	0.5840	0.0000	5.840e-1
$\gamma = 1.e - 12$		ITER=121		$\gamma = 1.e - 4$		ITER=66	

Figure 1. Example 1. $b(y) = 3\exp(-y^2)$.

From our point of view, in this case, the numerical results are qualitatively good in the interval I defined by the data η_{ij} , even under the presence of a relatively high level of noise (up to 10%). As one can expect, in the safety zone $I_w \setminus I$, the approximation becomes worse, but it seems still acceptable.

Example 2 We consider the recuperation of $a(x)b(y)$ with $a(x) = \sin(\pi x/4)$ and $b(y) = 3\exp(-y^2)$ from the data corresponding to the

source terms u_j given by (12) with (i_j, k_j) taking the values $(1, 1), (1, 8), (2, 2), (3, 5), (5, 6)$ and $(15, 4)$. No explicit noise is included here (i.e. $\delta = 0$). In Figure 2, we show the computed results for $\epsilon = 1.e - 12$ and $\varepsilon = 2.e - 9$, assuming $\hat{a}_1 \geq 1.e - 12$. This figure contains four graphics. On the top, the numerical approximation (obtained after 151 iterations) and the exact nonlinearity $a(x)b(y)$ appear on the left and on the right, respectively. On the bottom, the relative errors

$$\frac{|\check{a}(x_i)\check{b}(y_j) - a(x_i)b(y_j)|}{\max(1, |a(x_i)b(y_j)|)}$$

are shown.

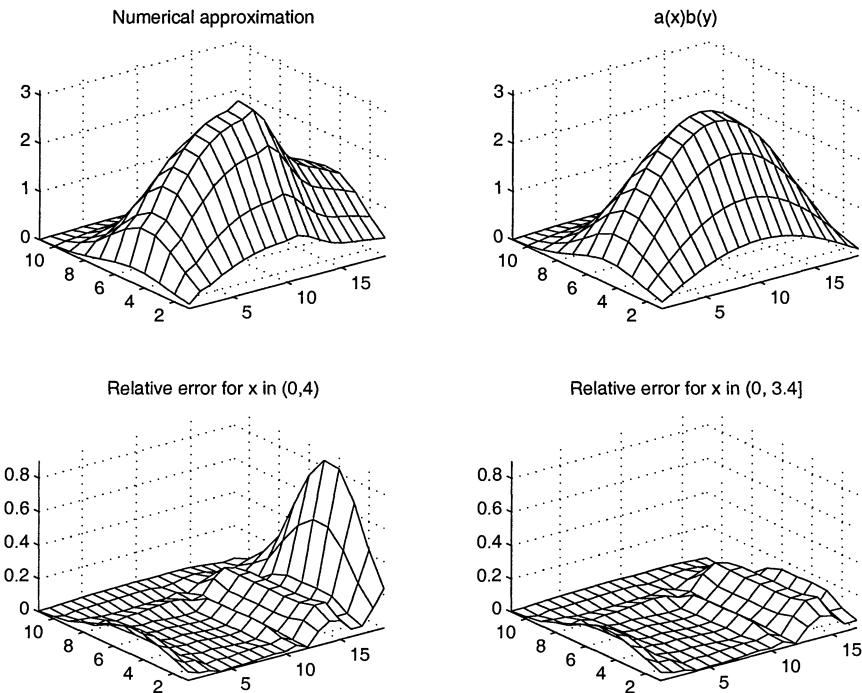


Figure 2. Example 2. $a(x) = \sin(\pi x/4)$ and $b(y) = 3\exp(-y^2)$.

In this example, we think that the computed results are relatively satisfactory when $x \in (0, 3.4]$ and they can be improved in $(3.4, 4)$. In our opinion, the deterioration in this zone can be due to the lack of information there: let us remind that ω_{in} and ω_{out} are both included in $[0, 2.1]$, while $\Omega = (0, 4)$. This difficulty does not appear when the nonlinear term is independent of the space variable x .

References

- [1] Fabre, C., Puel, J. P., and Zuazua, E. (1995). Approximate controllability of the semilinear heat equation. *Proc. Royal Soc. Edinburgh*, 125A:31–61.
- [2] Fernández, L. A. and Pola, C. (2001). Identification of a quasilinear parabolic equation from final data. *Int. J. Appl. Math. Comput. Sci.*, 11(4):859–879.
- [3] Fernández, L. A. and Zuazua, E. (1999). Approximate controllability for the semilinear heat equation involving gradient terms. *J. Optimization Theory & Appl.*, 101(2):307–328.
- [4] Isakov, V. (1998). *Inverse Problems for Partial Differential Equations*. Springer, New York-Berlin-Heidelberg.
- [5] Ladyzhenskaya, O. A., Solonnikov, V. A., and Ural'tseva, N. N. (1968). *Linear and Quasilinear Equations of Parabolic Type*. Am. Math. Soc, Providence, R. I.

ABOUT A GENERALIZED ALGEBRAIC RICCATI EQUATION

G. Freiling and A. Hochhaus

Department of Mathematics, University of Duisburg, D-47048 Duisburg, Germany

freiling@math.uni-duisburg.de, hochhaus@math.uni-duisburg.de

Abstract We describe the connection between a generalized algebraic Riccati equation and the corresponding generalized algebraic Riccati system and Kalman-Popov-Yakubovich system. Moreover, we present an iterative procedure for constructing its stabilizing solution.

Keywords: Generalized algebraic Riccati equation, generalized Lur'e system, generalized Kalman-Popov-Yakubovich system, stabilizing solution.

1. Introduction

The solution of stochastic linear quadratic optimal control problems with infinite time horizon leads to algebraic matrix equations of the form (see [3], [18])

$$A^*X + XA + Q + \Pi_1(X) - \\ - [S + XB + \Pi_{12}(X)][R + \Pi_2(X)]^+[S + XB + \Pi_{12}(X)]^* = 0, \quad (1.1)$$

where Z^+ is the Moore-Penrose inverse of a matrix Z and A, B, Q, R and S are given matrices of sizes $n \times n$, $n \times m$, $n \times n$, $m \times m$ and $n \times m$, respectively, such that

$$T := \begin{bmatrix} Q & S \\ S^* & R \end{bmatrix}$$

is hermitian. Moreover, the operator $\Pi: \mathcal{H}^n \rightarrow \mathcal{H}^{n+m}$ with

$$\Pi(X) := \begin{bmatrix} \Pi_1(X) & \Pi_{12}(X) \\ \Pi_{12}(X)^* & \Pi_2(X) \end{bmatrix}$$

is linear and positive, i.e. $X \geq 0$ implies $\Pi(X) \geq 0$. Here, \mathcal{H}^n stands for the real vector space of hermitian matrices of size n , and by $X \geq 0$ (or $X > 0$) it is denoted that $X = X^*$ is positive semidefinite (or positive definite).

If R is invertible, $S = 0$ and $\Pi \equiv 0$ then (1.1) reduces to the well-known algebraic Riccati equation

$$A^*X + XA + Q - XBR^{-1}B^*X = 0$$

which is of great importance in many fields of applied mathematics, e.g. optimal control theory. For that reason we call equation (1.1) *generalized algebraic Riccati equation*.

For $X \in \mathcal{H}^n$ we define the *dissipation matrix* associated with the generalized algebraic Riccati equation (1.1) as

$$\Lambda(X) := \begin{bmatrix} A^*X + XA + Q + \Pi_1(X) & S + XB + \Pi_{12}(X) \\ [S + XB + \Pi_{12}(X)]^* & R + \Pi_2(X) \end{bmatrix}. \quad (1.2)$$

Below we will describe how equation (1.1) is related to the (continuous-time) *generalized algebraic Riccati system* (or *Luré system*)

$$\Lambda(X) \begin{bmatrix} I \\ K \end{bmatrix} = 0 \quad (1.3)$$

in the unknowns $K \in \mathbb{C}^{m \times n}$, $X \in \mathcal{H}^n$ and to the *generalized Kalman-Popov-Yakubovich system*

$$\Lambda(X) = \begin{bmatrix} V^* \\ W^* \end{bmatrix} \begin{bmatrix} V & W \end{bmatrix} \quad (1.4)$$

in the unknowns $X \in \mathcal{H}^n$, $V \in \mathbb{C}^{m \times n}$ and $W \in \mathbb{C}^{m \times m}$.

For the case where R is invertible and $\Pi \equiv 0$ these results can be found in [17]. It turns out that under adequate additional restrictions (see Section 3) (1.1), (1.3) and (1.4) are equivalent.

Finally, in Section 4 we present an iterative procedure for computing the maximal (and stabilizing) solution X_+ of (1.1) which exists under certain generalized stabilizability and detectability assumptions; alternative algorithms for determining X_+ have been developed in [4] (via semidefinite programming associated with LMI's) and in [8], [14] (via Newton-type iteration).

Further results on (1.1) can be found in [2], [4], [5], [6], [7], [8], [9], [10] and [14].

2. Lyapunov equations and stability

In this section we present some preliminary results on the linearly perturbed algebraic Lyapunov equation

$$A^*X + XA + \Pi_1(X) + Q = 0, \quad (2.1)$$

where A and Q are given $n \times n$ matrices, Q is hermitian and $\Pi_1: \mathcal{H}^n \rightarrow \mathcal{H}^n$ is a positive linear operator. This equation plays a central role in the analysis of the generalized Riccati equation (1.1).

Define the continuous-time Lyapunov operator \mathcal{L}_A by

$$\mathcal{L}_A: \mathcal{H}^n \rightarrow \mathcal{H}^n, \quad X \mapsto A^*X + XA.$$

For an operator or a matrix T we denote by $\sigma(T)$ and $r(T)$ its spectrum and its spectral radius, respectively. The open left half-plane is \mathbb{C}_- .

The first theorem which can be found e.g. in [8] generalizes Lyapunov's stability theorem. A slightly modified version of this result appeared already in [12] (see also [11], Section III); in the case of time-varying coefficients a similar result has been proved in [9], Proposition 4.6.

2.1 Theorem. *The following statements are equivalent:*

- (i) $\sigma(A) \subset \mathbb{C}_-$ and $r(\mathcal{L}_A^{-1}\Pi_1) < 1$.
- (ii) $-(\mathcal{L}_A - \Pi_1)$ is inverse positive, i.e. $-(\mathcal{L}_A - \Pi_1)^{-1}$ exists and is a positive operator.
- (iii) There is some $X > 0$ such that $(\mathcal{L}_A + \Pi_1)(X) < 0$.
- (iv) For any $Q > 0$ equation (2.1) has a unique solution $X > 0$.
- (v) $\sigma(\mathcal{L}_A + \Pi_1) \subseteq \mathbb{C}_-$.

If any one of these conditions is fulfilled then A is called *c-stable relative to Π_1* .

It turns out that the classical definitions of stabilizability and detectability have to be replaced in our situation by the following generalizations.

2.2 Definition. A pair (A, B) of matrices $A \in \mathbb{C}^{n \times n}$ and $B \in \mathbb{C}^{n \times m}$ is said to be *c-stabilizable relative to Π* if there is a matrix F such that $A + BF$ is *c-stable relative to $\begin{bmatrix} I \\ F \end{bmatrix}^* \Pi \begin{bmatrix} I \\ F \end{bmatrix}$* .

According to Theorem 2.1 (A, B) is *c-stabilizable relative to Π* if and only if the inequality $(A + BF)^*X + X(A + BF) + \begin{bmatrix} I \\ F \end{bmatrix}^* \Pi(X) \begin{bmatrix} I \\ F \end{bmatrix} < 0$ is fulfilled by a pair (F, X) with $X > 0$.

Notice that the concept of mean-square stabilizability used in [12] is in the special case considered therein equivalent to *c-stabilizability relative to Π* .

2.3 Definition. A pair (C, A) of matrices $A \in \mathbb{C}^{n \times n}$ and $C \in \mathbb{C}^{m \times n}$ is said to be *c-detectable relative to Π_1* if there is a matrix $L \in \mathbb{C}^{n \times m}$ such that $A + LC$ is *c-stable relative to Π_1* .

Next we formulate a necessary condition for c -detectability which corresponds to the well-known Popov-Belevitch-Hautus criterion.

2.4 Lemma (see [14]). *If there exist a positive semidefinite matrix $V \neq 0$ with $CV = 0$ and some $\lambda \geq 0$ such that $(\mathcal{L}_A + \Pi_1)^{\text{adj}}(V) = \lambda V$, then (C, A) is not c -detectable relative to Π_1 .*

The following lemma generalizes results known from stability theory in the special case $\Pi_1 \equiv 0$.

2.5 Lemma (see [14]). *Suppose $Q \geq 0$ and (2.1) has a solution $X \geq 0$.*

- (i) *If $Q > 0$ then A is c -stable relative to Π_1 and we have $X > 0$.*
- (ii) *If (Q, A) is c -detectable relative to Π_1 then A is c -stable relative to Π_1 .*

3. Main results

For convenience of the reader we recall first the definition of the *Moore-Penrose inverse* and state some of its elementary properties which can be found for example in Section 20.5 of [16].

3.1 Definition. The *Moore-Penrose inverse* of an $m \times n$ matrix Z is the unique $n \times m$ matrix Z^+ satisfying the conditions

- (i) $Z^+ZZ^+ = Z^+$, $ZZ^+Z = Z$,
- (ii) $(Z^+Z)^* = Z^+Z$, $(ZZ^+)^* = ZZ^+$.

3.2 Lemma. *Let Z be an $m \times n$ matrix. Then:*

- (i) $(Z^+)^+ = Z$.
- (ii) $(Z^*)^+ = (Z^+)^*$.
- (iii) $(\lambda Z)^+ = \lambda^{-1}Z^+$ for all $\lambda \neq 0$.
- (iv) $\text{Ker } Z^+ = \text{Ker } Z^*$, $\text{Im } Z^+ = \text{Im } Z^*$.
- (v) *If Z is hermitian or positive semidefinite, then so is Z^+ .*
- (vi) $Z^+ = Z^*(ZZ^*)^+ = (Z^*Z)^+Z^*$.

The following lemma is taken from [1] (see Theorem 9.17 therein):

3.3 Lemma. *Assume that Z is an $m \times n$ matrix and W is a $p \times n$ matrix. Then the following statements are equivalent:*

- (i) $\text{Ker } Z \subseteq \text{Ker } W$.
- (ii) $W = WZ^+Z$.
- (iii) $W^+ = Z^+ZW^+$.

For every $X \in \mathcal{H}^n$ we introduce the corresponding *feedback matrix*

$$F(X) := -[R + \Pi_2(X)]^+[S + XB + \Pi_{12}(X)]^*.$$

The following lemma explains the relation between the generalized algebraic Riccati equation (1.1) and the generalized algebraic Riccati system (1.3).

3.4 Lemma. *A matrix $X \in \mathcal{H}^n$ is a solution of*

$$\begin{aligned} A^*X + XA + Q + \Pi_1(X) - [S + XB + \Pi_{12}(X)] \\ \times [R + \Pi_2(X)]^+ [S + XB + \Pi_{12}(X)]^* = 0 \end{aligned} \quad (3.1a)$$

with

$$\text{Ker}[R + \Pi_2(X)] \subseteq \text{Ker}[S + XB + \Pi_{12}(X)] \quad (3.1b)$$

if and only if there is a matrix $K \in \mathbb{C}^{m \times n}$ such that

$$\Lambda(X) \begin{bmatrix} I \\ K \end{bmatrix} = 0. \quad (3.2)$$

In this case $K = F(X)$, if additionally

$$\text{Ker}[R + \Pi_2(X)] \subseteq \text{Ker } K^*. \quad (3.3)$$

Proof. (i) If $X \in \mathcal{H}^n$ is a solution of (3.1a) and $K := F(X)$ is the corresponding feedback matrix, it is obvious that

$$A^*X + XA + Q + \Pi_1(X) + [S + XB + \Pi_{12}(X)]K = 0. \quad (3.2a)$$

Furthermore from Lemma 3.3 it follows that (3.1b) is equivalent to

$$-K^*[R + \Pi_2(X)] = S + XB + \Pi_{12}(X),$$

therefore

$$[S + XB + \Pi_{12}(X)]^* + [R + \Pi_2(X)]K = 0. \quad (3.2b)$$

Consequently K satisfies equation (3.2).

(ii) If $X \in \mathcal{H}^n$ and $K \in \mathbb{C}^{m \times n}$ are chosen such that (3.2b) holds, then it follows from the properties of the Moore-Penrose inverse that

$$\begin{aligned} [S + XB + \Pi_{12}(X)]K &= -K^*[R + \Pi_2(X)]K \\ &= -K^*[R + \Pi_2(X)][R + \Pi_2(X)]^+[R + \Pi_2(X)]K \\ &= -[S + XB + \Pi_{12}(X)][R + \Pi_2(X)]^+[S + XB + \Pi_{12}(X)]^*. \end{aligned}$$

Plugging this into (3.2a) yields that X satisfies (3.1a). Analogously it is obtained that

$$\begin{aligned} [S + XB + \Pi_{12}(X)][R + \Pi_2(X)]^+[R + \Pi_2(X)] \\ = -K^*[R + \Pi_2(X)][R + \Pi_2(X)]^+[R + \Pi_2(X)] \\ = -K^*[R + \Pi_2(X)] = S + XB + \Pi_{12}(X), \end{aligned}$$

and according to Lemma 3.3 this is equivalent to (3.1b).

(iii) Using parts (i) and (iv) of Lemma 3.2 and Lemma 3.3 it can easily be seen that relation (3.3) is equivalent to $\text{Ker}[R + \Pi_2(X)]^+ \subseteq \text{Ker } K^*$ and to $K = [R + \Pi_2(X)]^+[R + \Pi_2(X)]K$, respectively. Now premultiplication of (3.2b) with $[R + \Pi_2(X)]^+$ yields the last statement of the lemma. ■

The next theorem shows the relation between the generalized algebraic Riccati equation (1.1) and the generalized Kalman-Popov-Yakubovich system (1.4); notice that (1.4) provides a factorization of the dissipation matrix (1.2).

3.5 Theorem. *A matrix $X \in \mathcal{H}^n$ is a solution of*

$$\begin{aligned} A^*X + XA + Q + \Pi_1(X) - [S + XB + \Pi_{12}(X)] \\ \times [R + \Pi_2(X)]^+[S + XB + \Pi_{12}(X)]^* = 0 \end{aligned} \quad (3.4a)$$

with

$$\text{Ker}[R + \Pi_2(X)] \subseteq \text{Ker}[S + XB + \Pi_{12}(X)] \quad (3.4b)$$

$$\text{and} \quad R + \Pi_2(X) \geq 0, \quad (3.4c)$$

if and only there exist matrices $V \in \mathbb{C}^{m \times n}$ and $W \in \mathbb{C}^{m \times m}$ such that

$$\Lambda(X) = \begin{bmatrix} V^* \\ W^* \end{bmatrix} \begin{bmatrix} V & W \end{bmatrix} \quad (3.5a)$$

and

$$\text{Ker } W^* \subseteq \text{Ker } V^*. \quad (3.5b)$$

In this case $F(X) = -W^*V$.

Proof. (i) If $X \in \mathcal{H}^n$ is a solution of (3.4) then there exist a matrix $W \in \mathbb{C}^{m \times m}$ with $R + \Pi_2(X) = W^*W$ and (see Lemma 3.4) a matrix $K \in \mathbb{C}^{m \times n}$ such that (3.2) is fulfilled. Define $V := -WK$, then

$$[S + XB + \Pi_{12}(X)]^* = -[R + \Pi_2(X)]K = -W^*WK = W^*V$$

and, consequently,

$A^*X + XA + Q + \Pi_1(X) = -[S + XB + \Pi_{12}(X)]K = -V^*WK = V^*V$. Hence V and W satisfy (3.5a). From $V = -WK = -WW^*WK = WW^*V$ it follows with Lemma 3.3 that (3.5b) is valid.

(ii) Let $X \in \mathcal{H}^n$, $V \in \mathbb{C}^{m \times n}$ and $W \in \mathbb{C}^{m \times m}$ be chosen such that (3.5) holds. Then, according to Lemma 3.3, $V = WW^*V$. Defining $K := -W^*V$ it follows, using (3.5) again, that

$$[S + XB + \Pi_{12}(X)]K = -V^*WW^*V = -V^*V = -[A^*X + XA + Q + \Pi_1(X)]$$

and

$$[R + \Pi_2(X)]K = -W^*WW^*V = -W^*V = -[S + XB + \Pi_{12}(X)]^*.$$

Therefore X and K satisfy the algebraic Riccati system (3.2), moreover $R + \Pi_2(X) = W^*W \geq 0$. As a consequence of Lemma 3.4 X is a solution of the Kalman-Popov-Yakubovich system (3.5a). Finally, from Lemma 3.2 (vi) we obtain

$$-W^*V = -(W^*W)^+W^*V = -[R + \Pi_2(X)]^+[S + XB + \Pi_{12}(X)]^*,$$

which proves the statement of the theorem. ■

4. Computation of the stabilizing solution

For the formulation of the subsequent results we define $D(\mathcal{R})$ as the set of all $X \in \mathcal{H}^n$ such that

$$R + \Pi_2(X) \geq 0 \quad \text{and} \quad \text{Ker}[R + \Pi_2(X)] \subseteq \text{Ker}[S + XB + \Pi_{12}(X)]$$

and the generalized Riccati operator $\mathcal{R}: D(\mathcal{R}) \rightarrow \mathcal{H}^n$ by

$$\begin{aligned} \mathcal{R}(X) = & A^*X + XA + Q + \Pi_1(X) - [S + XB + \Pi_{12}(X)] \\ & \times [R + \Pi_2(X)]^+ [S + XB + \Pi_{12}(X)]^*. \end{aligned}$$

We agree that all statements concerning solutions X of $\mathcal{R}(X) = 0$ are made under the additional hypothesis $X \in D(\mathcal{R})$.

If $X \in \mathcal{H}^n$ is a solution of $\mathcal{R}(X) = 0$ and $F = F(X)$ denotes the corresponding feedback matrix then X is called *stabilizing* (resp. *almost stabilizing*) if $\sigma(\mathcal{L}_{A+BF} + \hat{\Pi})$ with $\hat{\Pi} = [\begin{smallmatrix} I \\ F \end{smallmatrix}]^* \Pi [\begin{smallmatrix} I \\ F \end{smallmatrix}]$ is contained in the open (resp. closed) left half-plane.

In the following we propose an algorithm for the computation of the stabilizing solution X_+ of $\mathcal{R}(X) = 0$ which is based on Theorem 3.5. For convenience below we use the following assumptions which ensure the existence of X_+ :

- (H1) (A, B) ist c -stabilizable relative to Π ,
- (H2) $R > 0$, $T := (\begin{smallmatrix} Q & S \\ S^* & R \end{smallmatrix}) \geq 0$,
- (H3) $(Q - SR^{-1}S^*, A - BR^{-1}S^*)$ ist c -detectable relative to $[\begin{smallmatrix} I \\ -R^{-1}S^* \end{smallmatrix}]^* \Pi [\begin{smallmatrix} I \\ -R^{-1}S^* \end{smallmatrix}]$.

The next two lemmata have been proved in [14]:

4.1 Lemma. *If $\mathcal{R}(X) = 0$ has a stabilizing solution X_+ , then $X_+ \geq X$ for every solution X of $\mathcal{R}(X) \geq 0$. In particular, X_+ is the (unique) maximal solution of $\mathcal{R}(X) = 0$.*

4.2 Lemma. *Assume that the hypotheses (H2) and (H3) hold. Then every positive semidefinite solution of $\mathcal{R}(X) = 0$ is stabilizing.*

By hypothesis (H1), there is an F_0 such that $A_0 := A + BF_0$ is c -stable relative to $[\begin{smallmatrix} I \\ F_0 \end{smallmatrix}]^* \Pi [\begin{smallmatrix} I \\ F_0 \end{smallmatrix}]$. Therefore, it follows from Theorem 2.1 that the linearly perturbed Lyapunov equation

$$A_0^*X_1 + X_1A_0 + \left[\begin{array}{c|c} I & \\ \hline F_0 & \end{array} \right]^* [T + \Pi(X_1)] \left[\begin{array}{c|c} I & \\ \hline F_0 & \end{array} \right] = 0$$

has a unique solution X_1 which is positiv semidefinite since $T \geq 0$. Hence $R + \Pi_2(X_1) > 0$, and consequently there exist matrices $V_1 \in$

$\mathbb{C}^{m \times n}$ and $W_1 \in GL(m, \mathbb{C})$ with $R + \Pi_2(X_1) = W_1^* W_1$ and $V_1^* W_1 = S + X_1 B + \Pi_{12}(X_1)$. We use induction to construct sequences of matrices $\{A_i\}_{i=0}^\infty$, $\{F_i\}_{i=0}^\infty$, $\{V_i\}_{i=1}^\infty$, $\{W_i\}_{i=1}^\infty$ and $\{X_i\}_{i=1}^\infty$ with certain properties. Thus, assume that for some $m \geq 1$ we have already determined matrices $\{A_i\}_{i=0}^{m-1}$, $\{F_i\}_{i=0}^{m-1}$, $\{V_i\}_{i=1}^m$, $\{W_i\}_{i=1}^m$ and $\{X_i\}_{i=1}^m$ with

$$X_1 \geq X_2 \geq \dots \geq X_m \geq 0, \quad A_i = A + BF_i, \quad i = 0, 1, \dots, m-1,$$

where

$$W_i^* W_i = R + \Pi_2(X_i), \quad V_i^* W_i = S + X_i B + \Pi_{12}(X_i)$$

for $i = 1, 2, \dots, m$,

$$\begin{aligned} F_i &= -W_i^{-1} V_i, \quad i = 1, 2, \dots, m-1, \\ A_i^* X_{i+1} + X_{i+1} A_i + \begin{bmatrix} I \\ F_i \end{bmatrix}^* [T + \Pi(X_{i+1})] \begin{bmatrix} I \\ F_i \end{bmatrix} &= 0 \end{aligned} \quad (4.1)$$

for $i = 0, 1, \dots, m-1$ and the matrices A_i are c -stable relative to $\begin{bmatrix} I \\ F_i \end{bmatrix}^* \Pi \begin{bmatrix} I \\ F_i \end{bmatrix}$, $i = 0, 1, \dots, m-1$. Now define $F_m := -W_m^{-1} V_m$ and $A_m := A + BF_m$. Then

$$\begin{aligned} A_m^* X_m + X_m A_m + \begin{bmatrix} I \\ F_m \end{bmatrix}^* [T + \Pi(X_m)] \begin{bmatrix} I \\ F_m \end{bmatrix} \\ + (F_m - F_{m-1})^* [R + \Pi_2(X_m)] (F_m - F_{m-1}) &= 0, \end{aligned}$$

and as in the proof of Lemma 4.2 (see [14]) it follows with the hypotheses (H2) and (H3) that A_m is c -stable relative to $\begin{bmatrix} I \\ F_m \end{bmatrix}^* \Pi \begin{bmatrix} I \\ F_m \end{bmatrix}$. According to Theorem 2.1 the linearly perturbed Lyapunov equation

$$A_m^* X_{m+1} + X_{m+1} A_m + \begin{bmatrix} I \\ F_m \end{bmatrix}^* [T + \Pi(X_{m+1})] \begin{bmatrix} I \\ F_m \end{bmatrix} = 0$$

has a unique solution $X_{m+1} \geq 0$, and the difference $X_m - X_{m+1}$ satisfies

$$\begin{aligned} A_m^* (X_m - X_{m+1}) + (X_m - X_{m+1}) A_m + \begin{bmatrix} I \\ F_m \end{bmatrix}^* \Pi(X_m - X_{m+1}) \begin{bmatrix} I \\ F_m \end{bmatrix} \\ + (F_m - F_{m-1})^* [R + \Pi_2(X_m)] (F_m - F_{m-1}) &= 0 \end{aligned}$$

(see the proof of Theorem 5.2 in [14]). Since A_m is c -stable relative to $\begin{bmatrix} I \\ F_m \end{bmatrix}^* \Pi \begin{bmatrix} I \\ F_m \end{bmatrix}$, it follows from Theorem 2.1 that $X_m \geq X_{m+1}$.

So $\{X_i\}_{i=1}^\infty$ is a nonincreasing sequence of positive semidefinite matrices. Hence the limit $X_+ := \lim_{i \rightarrow \infty} X_i$ exists, and $X_+ \geq 0$. Furthermore, X_+ satisfies $\mathcal{R}(X) = 0$. According to Lemma 4.2 X_+ is stabilizing and by Lemma 4.1 also maximal.

A discrete-time version of the algorithm presented above could also be developed (for the basic theory see [13]); in the case $\Pi \equiv 0$ it can be found in [15].

4.3 Remark. Note that the maximal solution X_+ also exists (see [14]) if the hypothesis (H1) holds and if (H2) and (H3) are replaced by

(H4) There is a matrix $\hat{X} \in D(\mathcal{R})$ with $\text{Ker}[R + \Pi_2(\hat{X})] \subseteq \text{Ker } B$ for which $\mathcal{R}(\hat{X}) \geq 0$.

Under these weaker assumptions X_+ is at least almost stabilizing and can be computed by replacing (4.1) in the preceding algorithm by

$$A_i^* X_{i+1} + X_{i+1} A_i + \begin{bmatrix} I \\ F_i \end{bmatrix}^* [T + \Pi(X_{i+1})] \begin{bmatrix} I \\ F_i \end{bmatrix} + \frac{1}{i+1} I = 0;$$

or, equivalently, by applying the modified Newton algorithm presented in [14].

Using the theory of the differential equation $-\dot{X} = \mathcal{R}(X)$, developed in [14], we obtain in an elementary way another existence result for $\mathcal{R}(X) = 0$.

4.4 Lemma. $\mathcal{R}(X) = 0$ has a solution \hat{X} with $\Pi_2(\hat{X}) > -R$ if and only if there exist $X_1 \geq X_2$ with $\Pi_2(X_2) > -R$ and

$$\mathcal{R}(X_1) \leq 0 \leq \mathcal{R}(X_2). \quad (4.2)$$

Proof. If \hat{X} is a solution of $\mathcal{R}(X) = 0$ with $\Pi_2(\hat{X}) > -R$ then (4.2) holds trivially.

Conversely, it follows from the Monotonicity Lemma 6.1 and the Comparison Theorem 4.5 in [14] that the solutions $X(\cdot, X_1)$ resp. $X(\cdot, X_2)$ of $-\dot{X} = \mathcal{R}(X)$ with $X(0, X_i) = X_i$, $i = 1, 2$, are decreasing resp. increasing as t is decreasing; moreover

$$X_2 \leq X(t, X_2) \leq X(t, X_1) \leq X_1 \quad \text{for all } t \leq 0.$$

Hence the limits

$$X_{2,\infty} = \lim_{t \rightarrow -\infty} X(t, X_2) \leq X_{1,\infty} = \lim_{t \rightarrow -\infty} X(t, X_1)$$

exist and solve $\mathcal{R}(X) = 0$. Notice that $X(t, X_0)$ exists for $t \leq 0$ if $X_2 \leq X_0 \leq X_1$. ■

References

- [1] Ahlbrandt, C. D.; Peterson, A. C.: Discrete Hamiltonian systems, *Kluwer Academic Publishers Group, Dordrecht*, 1996.
- [2] Ait Rami, M.; Chen, X.; Moore, J. B.; Zhou, X. Y.: Solvability and asymptotic behavior of generalized Riccati equations arising in indefinite stochastic LQ controls, *IEEE Trans. Automat. Control* **46** (2001), 428–440.

- [3] Ait Rami, M.; Moore, J. B.; Zhou, X. Y.: Indefinite stochastic linear quadratic control and generalized differential Riccati equations, Preprint.
- [4] Ait Rami, M.; Zhou, X. Y.: Linear matrix inequalities, Riccati equations, and indefinite stochastic linear quadratic controls, *IEEE Trans. Automat. Control* **45** (2000), 1131–1143.
- [5] Chen, S.; Li, X.; Zhou, X. Y.: Stochastic linear quadratic regulators with indefinite control weight costs, *SIAM J. Control Optim.* **36** (1998), 1685–1702.
- [6] Chen, S.; Yong, J.: Stochastic linear quadratic optimal control problems, *Appl. Math. Optim.* **43** (2001), 21–45.
- [7] Chen, S.; Zhou, X. Y.: Stochastic linear quadratic regulators with indefinite control weight costs II, *SIAM J. Control Optim.* **39** (2000), 1065–1081.
- [8] Damm, T.; Hinrichsen, D.: Newton's method for a rational matrix equation occurring in stochastic control, *Linear Algebra Appl.* **332/334** (2001), 81–109.
- [9] Dragan, V.; Morozan, T.: Stability and robust stabilization to linear stochastic systems described by differential equations with markovian jumping and multiplicative white noise, *Stochastic Analysis*, **20**, no.1, 2002, 33-92.
- [10] Dragan, V.; Morozan, T.: Systems of matrix rational differential equations arising in connection with linear stochastic systems with markovian jumping, *Preprint No. 9/2000, Institutul de Matematică al Academiei Române* (2000); to appear in *J. Differential Equ.*, 2003.
- [11] Feng, X.; Loparo, K. A.; Ji, Y.; Chizeck, H. J.: Stochastic stability properties of jump linear systems, *IEEE Trans. Automat. Control* **37** (1992), 38–53.
- [12] Fragoso, M. D.; Costa, O. L. V.; de Souza, C. E.: A new approach to linearly perturbed Riccati equations arising in stochastic control, *Appl. Math. Optim.* **37** (1998), 99–126.
- [13] Freiling, G.; Hochhaus, A.: Properties of the solutions of rational matrix difference equations. Advances in difference equations, IV, *Comput. Math. Appl.*, to appear.
- [14] Freiling, G.; Hochhaus, A.: On a class of rational matrix differential equations arising in stochastic control, *Linear Algebra Appl.*, to appear.
- [15] Halanay, A.; Samuel, J.: Differential equations, discrete systems and control, *Kluwer Academic Publishers, Dordrecht*, 1997.
- [16] Harville, D. A.: Matrix algebra from a statistician's perspective, *Springer-Verlag, New York*, 1997.
- [17] Ionescu, V.; Oară, C.; Weiss, M.: Generalized Riccati theory and robust control, *John Wiley & Sons, Ltd., Chichester*, 1999.
- [18] Yong, J.; Zhou, X. Y.: Stochastic controls, *Springer-Verlag, New York*, 1999.

ON THE DYNAMICS OF CONTRACTING RELATIONS

Vasile Glavan

Moldova State University

Chișinău, Republic of Moldova

University of Podlasie

Siedlce, Poland

glavan@usm.md vgjavan@ap.siedlce.pl

Valeriu Guțu

Moldova State University

60, A. Mateevici Str.

MD-2009 Chișinău, Republic of Moldova

gutu@usm.md

Abstract Attractors of pointwise bounded and closed relations in metric spaces are considered. An analog of the asymptotic phase theorem for the contracting relations has been proved. The Shadowing Property near the attractor of a contracting relation is stated. As a consequence it is proved that the periodic points form a dense subset of the attractor.

Keywords: Relations, Attractors, Shadowing Property

Introduction

Dynamical systems have their roots in the pioneering works of Poincaré on the qualitative theory of differential equations. Discrete dynamical systems as iteration of diffeomorphisms appeared firstly as section maps of flows near periodic orbits. Later the concept evolved until the iteration of diffeomorphisms, or, recently, of noninvertible maps, became an independent subject of researches. More recently, iteration of finite collections of contractions, named as Iterated Function Systems (IFS), became the main tool for studying of the geometry of fractals as attractors, as well as the chaotic dynamics on these attractors. Until now the IFS remain one of the mostly elaborated multivalued dynamical systems.

On the other hand, infinite dimensional dynamical systems, generated by partial differential equations without uniqueness (see [1], [2]), as well as control systems (see [3]), gave new motivations for studying the set-valued dynamics or, more precisely, the iterations of relations.

Evidently, such setting seems to be hopelessly general. However, it is demonstrated in [4] and [5] that the concept of limit sets, attractors, orbit and pseudo-orbit, transitivity and recurrence, known for ordinary flows, naturally generalizes for relations too (at least in the compact phase space setting).

The main goal of this paper consists in the developing of the notion of attractor for a relation in the noncompact setting, and in the studying of the dynamics near this attractor. One of the mostly used methods is the shadowing of pseudo-orbits. We give a criterion for a closed and bounded set to be the attractor of a contracting relation and prove the Shadowing Property of a such relation near the attractor. As a consequence, we obtain that the periodic points form a dense subset of the attractor.

1. Dynamics of relations

Let (X, d) be a complete metric space. A *relation on X* is a subset $f \subset X \times X$. Any relation can be regarded as a (set-valued) function from X to the power set $\mathcal{P}(X)$, associating to each $x \in X$ a subset $f(x)$ of X . These two aspects of relations (set theoretical and functional) allows one to apply subset operations, such as union, intersection and closure, on the one hand, and the functional operations, such as composition, inverse, identity, on the other hand.

The neighborhoods of the diagonal as $V_\varepsilon = \{(x_1, x_2) \in X \times X : d(x_1, x_2) < \varepsilon\}$, and similarly for \bar{V}_ε with nonstrict inequality, are important examples of relations.

A relation on a metric space is said to be *closed*, if it is a closed subset of the Cartesian product of the space with itself. In a compact space this is equivalent to the upper semicontinuity of the relation (see [4]).

For two relations $f, g : X \rightarrow \mathcal{P}(X)$ we define the *composition* $g \circ f : X \rightarrow \mathcal{P}(X)$ by: $(x, y) \in g \circ f$, if there exists $z \in X$ such that $(x, z) \in f$ and $(z, y) \in g$. The *inverse* of f is, by definition, $f^{-1} := \{(y, x) : (x, y) \in f\}$. The composition is associative, so for $n \in \mathbb{N}$ we define f^n to be the n -fold composition of f , and similarly $f^{-n} := (f^{-1})^n$. From associativity it follows that $f^{m+n} = f^m \circ f^n$ for $m, n \geq 0$ or $m, n \leq 0$. We call $\mathcal{O}f := \bigcup_{n \geq 1} f^n$ the *orbit relation*.

A subset $A \subset X$ is called *positive invariant* with respect to a relation f on X (written “ A is $f +$ invariant”), if $f[A] \subset A$, where $f[A] = \bigcup_{a \in A} f(a)$.

Further, A is called *f invariant*, if $f[A] = A$. The last means that A is f -invariant and, in addition, $f^{-1}(x) \cap A \neq \emptyset$ for all $x \in A$.

For relations the analogous of the orbit is the notion of the chain. The finite or infinite sequence $\{x_n\} \subset X$ is called a *chain* for a relation $f : X \rightarrow \mathcal{P}(X)$, if $x_{n+1} \in f(x_n)$ for all n , or, in other words, if $(x_n, x_{n+1}) \in f$.

Given a relation $f : X \rightarrow \mathcal{P}(X)$, a point $x \in X$ is called a *fixed point* for f , if $x \in f(x)$. Thus, $x \in Of(x)$, if and only if there exists $n \geq 1$ such that $x \in f^n(x)$. Such point is called a *periodic point* for f .

Recall that for any two closed and bounded subsets B_1 and B_2 of a metric space (X, d) the quantity $\varrho(B_1, B_2)$, given by

$$\varrho(B_1, B_2) := \sup_{b_1 \in B_1} \inf_{b_2 \in B_2} d(b_1, b_2),$$

defines the Hausdorff-Pompeiu metric as follows:

$$H(B_1, B_2) := \max\{\varrho(B_1, B_2), \varrho(B_2, B_1)\}.$$

This metric on the space $\mathcal{P}_{b,cl}(X)$ of all nonempty bounded and closed subsets of X is complete, if d is complete.

The following lemmas will be used in the next proofs.

Lemma 1.1. *For any nonempty bounded subsets $B_1, B_2 \subset X$ and a point $x \in X$ the following inequality holds: $\varrho(x, B_1) \leq \varrho(x, B_2) + H(B_1, B_2)$.*

Proof. In the inequality $d(x, b_1) \leq d(x, b_2) + d(b_1, b_2)$, $(x, b_1, b_2 \in X)$, fix $b_2 \in B_2$ and take infimum on B_1

$$\inf_{b_1 \in B_1} d(x, b_1) \leq d(x, b_2) + \inf_{b_1 \in B_1} d(b_1, b_2).$$

The last term is not greater than $H(B_1, B_2)$. So we have for any $b_2 \in B_2$:

$$\varrho(x, B_1) \leq d(x, b_2) + H(B_1, B_2).$$

Taking infimum on B_2 , we obtain the desired inequality. ■

Lemma 1.2. *Let $f : X \rightarrow \mathcal{P}_{b,cl}(X)$ be a relation such that*

$$H(f(x), f(y)) \leq \lambda d(x, y) \quad (x, y \in X)$$

for some $\lambda > 0$. Then $H(f[B_1], f[B_2]) \leq \lambda H(B_1, B_2)$ for any nonempty bounded subsets $B_1, B_2 \subset X$.

Proof. Given any subsets $B_1, B_2 \subset X$, we have

$$H(f(b_1), f(b_2)) \leq \lambda d(b_1, b_2) \quad (b_1 \in B_1, b_2 \in B_2).$$

Passing consecutively to infimum on B_2 and supremum on B_1 , we obtain $\varrho(f[B_1], f[B_2]) \leq \lambda \varrho(B_1, B_2)$. Analogously, $\varrho(f[B_2], f[B_1]) \leq \lambda \varrho(B_2, B_1)$. Thus, $H(f[B_1], f[B_2]) \leq \lambda H(B_1, B_2)$. ■

Definition 1.1. A relation $f : X \rightarrow \mathcal{P}_{b,cl}(X)$ such that

$$H(f(x), f(y)) \leq \lambda d(x, y) \quad (x, y \in X)$$

for some $0 < \lambda < 1$ is called a *contracting relation* with the ratio λ .

Remark 1.1. Hyperbolic IFS and IFS with condensation (see [6]) consist one of the mostly developed examples of contracting relations.

2. Attractors in relations

The dynamics of a contracting mapping is trivial: the unique fixed point attracts all the points and even all bounded subsets of the phase space. As for set-valued functions the dynamics is much more complicated. Firstly, a fixed point, if it exists, need not be unique, nor attractive. Moreover, the set of fixed points is not even invariant. Secondly, the attractor of a hyperbolic IFS, a particular case of a contracting relation, is an invariant set, although containing the fixed points, it is a scene of a much more complicated dynamics, e.g. it contains a dense subset of periodic points as well as a dense chain [6].

There are various definitions of attractor. In ordinary dynamics (e.g. iterations of mappings) by an attractor one usually means an invariant set, which is dynamically indivisible and whose basin – the set of attracted points – is a large set. The dynamical indivisibility is understood as the existence of a dense orbit. As for the basin, it must contain a neighborhood of the attractor, or at least the nonvoid interior, sometimes positive Lebesgue measure is required. At the same time, relations need not be continuous or even semicontinuous, so for the basin to contain an (open) neighborhood is not an adequate demand. The “dynamical indivisibility” in the set-valued setting is also not very well understood.

In the case of compact spaces in [4] (see also [5]) the following definition has been proposed: A is an attractor, if it is invariant and there exists a closed neighborhood V of A such that $\bigcap_{n \geq 0} f^n[V]$ is contained in A

In [1] and [2] the invariance $f[A] = A$ is relaxed up to the condition $f[A] \supset A$ with the assumption that A attracts any bounded subset of a neighborhood of A .

Definition 2.1. Let X be a metric space and $f : X \rightarrow \mathcal{P}(X)$ be a closed relation. A closed subset $A \subset X$ is called an *attractor* for f , if:

- i) $f[A] \supset A$;
- ii) there is a closed neighborhood V of A such that $\bigcap_{n \geq 0} f^n[V] \subset A$.

Remark 2.1. The second inclusion is, in fact, an equality. For, observe that $A \subset V$ and $A \subset f[A]$ imply the inclusions $A \subset f^n[V]$ for all $n \in \mathbb{N}$ and, hence, $A \subset \bigcap_{n \geq 0} f^n[V]$.

Remark 2.2. Analogously, the first inclusion actually is an equality. For, use the second equality and the inclusions

$$\bigcap_{n \geq 0} f^n[V] = A \subset f[A] = f\left(\bigcap_{n \geq 0} f^n[V]\right) = \bigcap_{n \geq 1} f^n[V] \subset A.$$

Remark 2.3. For the “Cantor’s relation” $f : \mathbf{R} \rightarrow \mathbf{R}$, $f(x) = \{\frac{1}{3}x\} \cup \{\frac{1}{3}x + \frac{2}{3}\}$ the set of fixed points $A = \{0; 1\}$ satisfies the inclusion $A \subset f[A]$, but the second condition does not hold.

Theorem 2.1. *For any contracting relation on a complete metric space there exists an unique bounded and closed invariant set.*

Proof. Let f be a contracting relation on X . Denote by $f_* : \mathcal{P}_{b,cl}(X) \rightarrow \mathcal{P}_{b,cl}(X)$, $f_*(A) = f[A]$, the corresponding map on the complete metric space $\mathcal{P}_{b,cl}(X)$ of nonempty bounded and closed subsets of X , endowed with the Hausdorff-Pompeiu metric H . By Lemma 1.2, f_* is a Lipschitz mapping, having the same contraction ratio as f . Hence, there exists an unique fixed point $A \in \mathcal{P}_{b,cl}(X)$. This means that the equation $f[A] = A$ has a solution in $\mathcal{P}_{b,cl}(X)$ and this solution is unique. ■

In the case of a hyperbolic IFS Barnsley [6] calls a set A , such that $f[A] = A$, “the attractor” for IFS. The following theorem shows that in the case of a contracting relation (and of a hyperbolic IFS, as a particular case) the Barnsley’s notion of attractor coincides with Definition 2.1.

Theorem 2.2. *A bounded and closed subset $A \subset X$ is an attractor for a contracting relation f , if and only if A is a compact invariant set for f .*

Proof. If A is a bounded attractor for f , then A is a bounded invariant set for f (see Remark 2.2). So A is an unique fixed point for f_* in the metric space $\mathcal{P}_{b,cl}(X)$. But we can repeat the proof of the Theorem 2.1 for the metric space $\mathcal{P}_{cp}(X)$ of nonempty compact subsets of X and to obtain an unique solution $A \in \mathcal{P}_{cp}(X)$ of the equation $f[A] = A$. Since any compact subset is also bounded and closed, the bounded and closed invariant set, given by Theorem 2.1, is actually compact.

Conversely, assume that A is a compact, and so, a bounded and closed invariant set for f . Take a closed neighborhood V of A of small enough radius. It is bounded as well.

By Lemma 1.2, we have for any $n \in \mathbf{N}$

$$\varrho(f^n[V], A) = \varrho(f^n[V], f^n[A]) \leq \lambda^n \varrho(V, A).$$

Because of the inclusion $f^{n+1}[V] \subset f^n[V]$, the last inequality implies:

$$\varrho\left(\bigcap_{k=0}^n f^k[V], A\right) \leq \lambda^n \varrho(V, A) \quad (n \in \mathbf{N}).$$

Passing to the limit as $n \rightarrow \infty$, we obtain $\varrho\left(\bigcap_{k \geq 0} f^k[V], A\right) = 0$, which is equivalent to the inclusion $\bigcap_{k \geq 0} f^k[V] \subset A$. Hence, A is an attractor. ■

Example. Let $\mathcal{C}[0, 1]$ be the space of continuous functions on $[0, 1]$ with the sup-metric. Fix $\psi \in \mathcal{C}[0, 1]$ and consider the relation $f : \varphi \mapsto \{\frac{1}{3}\varphi\} \cup \{\frac{1}{3}\varphi + \frac{2}{3}\psi\}$. Using Theorem 2.2, it is easy to check that the compact subset $K \subset \mathcal{C}[0, 1]$, $K = \{\tau \cdot \psi : \tau \in C\}$, where C stands for the middle-third Cantor set, is the attractor of this relation.

Theorem 2.3. *Let $f : X \rightarrow \mathcal{P}_{cp}(X)$ be a compact valued contracting relation with the ratio $0 < \lambda < 1$. Then for any chain $\{x_n\} \subset X$ and any $y_0 \in X$ there exists a chain $\{y_n\} \subset X$, starting in y_0 , such that*

$$d(x_n, y_n) \leq \lambda^n d(x_0, y_0) \quad (n \in \mathbb{N}).$$

Proof. It is known that for any nonempty compact subsets $A, B \subset X$ and any $a \in A$ there exists $b \in B$ such that $d(a, b) \leq H(A, B)$ (see, e.g. [7]). Therefore, for compact subsets $f(x_0), f(y_0) \subset X$ and given $x_1 \in f(x_0)$ there exists $y_1 \in f(y_0)$ such that

$$d(x_1, y_1) \leq H(f(x_0), f(y_0)) \leq \lambda d(x_0, y_0).$$

Similarly, for given $x_2 \in f(x_1)$ there exists $y_2 \in f(y_1)$ such that

$$d(x_2, y_2) \leq H(f(x_1), f(y_1)) \leq \lambda d(x_1, y_1) \leq \lambda^2 d(x_0, y_0).$$

Moreover, inductively, for any $n \in \mathbb{N}$ there exists $y_n \in f(y_{n-1})$ such that

$$d(x_n, y_n) \leq \lambda^n d(x_0, y_0).$$

The chain $\{y_n\}$ is a required one. ■

As a corollary we obtain immediately the following result.

Theorem 2.4 (Asymptotic phase theorem for contracting relations). *Let $f : X \rightarrow \mathcal{P}_{cp}(X)$ be a compact valued contracting relation with the ratio $0 < \lambda < 1$ and A stand for the attractor of f . Then for any chain $\{x_n\} \subset X$ and any $a_0 \in A$ there exists a chain $\{a_n\} \subset A$, starting in a_0 , such that $d(x_n, a_n) \leq \lambda^n d(x_0, a_0)$ ($n \in \mathbb{N}$).*

3. Skew-product relations

This section is devoted to fiberwise contracting skew-product relations. We prove that any almost invariant cross-section can be approximated by an invariant one. This abstract setting will be applied in the next section for proving one of the main results of the paper – the Shadowing Property of contracting relations.

Given a relation $f : X \rightarrow \mathcal{P}(X)$, a set Y and an injective mapping $\tau : Y \rightarrow Y$, let $F = (\tau, f) : Y \times X \rightarrow \mathcal{P}(Y \times X)$ be the corresponding relation on $Y \times X$. With $p : Y \times X \rightarrow X$ for the projection, one has the following commutative diagram of relations:

$$\begin{array}{ccc} Y \times X & \xrightarrow{F} & Y \times X \\ p \downarrow & & \downarrow p \\ Y & \xrightarrow{\tau} & Y \end{array}$$

Commutativity means the equality $\tau \circ p = p \circ F$. A function $\sigma : Y \rightarrow Y \times X$ such that $p \circ \sigma = \text{id}_Y$ is called a *cross-section*. It is clear that any cross-section is of the form $y \mapsto (y, \varphi(y))$ for some function $\varphi : Y \rightarrow X$. In what follows we use the same notation for the cross-section and for this function.

Definition 3.1. One says that the cross-section $\sigma : Y \rightarrow Y \times X$ is *F +invariant*, if $\sigma \circ \tau(y) \in F \circ \sigma(y)$ for any $y \in Y$.

Definition 3.2. Given $\varepsilon > 0$, we call the cross-section σ as (F, ε) +invariant, if $\sigma \circ \tau(y) \in \overline{V_\varepsilon} \circ (F \circ \sigma(y))$, or, in other words, if $\varrho(\sigma \circ \tau(y), F \circ \sigma(y)) \leq \varepsilon$ for any $y \in Y$.

Remark 3.1. If f is pointwise closed and bounded, then the $(F, 0)$ +invariance coincides with F +invariance, which, in turn, means that for any $y \in Y$ the point $\sigma \circ \tau(y)$ belongs to the subset $F \circ \sigma(y)$. Similarly, the (F, ε) +invariance means that the point $\sigma \circ \tau(y)$ is ε -close to the subset $F \circ \sigma(y)$, which, in turn, is equivalent to the +invariance with respect to the relation $(\tau, \overline{V_\varepsilon} \circ f) : Y \times X \rightarrow \mathcal{P}(Y \times X)$.

Theorem 3.1. Let (X, d) be a complete metric space and $f : X \rightarrow \mathcal{P}_{b,cl}(X)$ be a relation on X . Let Y be a set and $\tau : Y \rightarrow Y$ be an injection. Assume that f is a contraction with ratio $0 < \lambda < 1$. Then for any $\varepsilon > 0$ there is $\delta > 0$ such that for any (F, δ) +invariant cross-section $\psi : Y \rightarrow Y \times X$ there exists a F +invariant cross-section $\varphi : Y \rightarrow Y \times X$ such that $\bar{d}(\varphi, \psi) := \sup\{d(\varphi(y), \psi(y)) : y \in Y\} \leq \varepsilon$.

Proof. Fix $\varepsilon > 0$ and take $\delta = (1 - \lambda)\varepsilon$. Let ψ be any (F, δ) +invariant cross-section. Define X_ψ as the space of all cross-sections $\alpha : Y \rightarrow Y \times X$ such that $\bar{d}(\psi, \alpha) < \infty$. It is easy to check that (X_ψ, \bar{d}) is a complete metric space.

Let $X_{\psi, \varepsilon}$ denote the closed ball of radius ε with center ψ in X_ψ . Define a set-valued function (a relation) Φ on $X_{\psi, \varepsilon}$ as follows: $\Phi(\alpha) = F \circ \alpha \circ \tau^{-1} (\alpha \in X_{\psi, \varepsilon})$. Each value of this function is a closed and bounded subset of X_ψ . For, estimate

$$\text{diam}\Phi(\alpha) \leq \bar{H}(\Phi(\alpha), \Phi(\psi)) \leq \lambda \bar{d}(\alpha, \psi) \leq \lambda \varepsilon,$$

so $\Phi(\alpha)$ is a bounded set. Here \bar{H} stands for the Hausdorff-Pompeiu metric on $\mathcal{P}_{b,cl}(X_\psi)$.

To prove the closedness of $\Phi(\alpha)$ take a convergent sequence $\{\xi_n\} \subset \Phi(\alpha)$ with $\xi = \lim_{n \rightarrow \infty} \xi_n$. One has $\xi_n(y) \in f \circ \alpha \circ \tau^{-1}(y)$ ($y \in Y$). Since

$f \circ \alpha \circ \tau^{-1}(y)$ is a closed subset of X , we obtain that $\xi(y) \in f \circ \alpha \circ \tau^{-1}(y)$ also, which implies $\xi \in \Phi(\alpha)$.

Moreover, for any $\alpha \in X_{\psi,\varepsilon}$ we have that $\Phi(\alpha) \cap X_{\psi,\varepsilon}$ is a closed nonempty subset of $X_{\psi,\varepsilon}$. To show this it is enough to state the inequality $\bar{\varrho}(\psi, \Phi(\alpha)) \leq \varepsilon$ ($\bar{\varrho}$ is defined similarly as \bar{H}).

Fix $y \in Y$. For the taken value $\delta = (1 - \lambda)\varepsilon$, using Lemma 1.1, we obtain:

$$\begin{aligned}\varrho(\psi(y), F \circ \alpha \circ \tau^{-1}(y)) &\leq \varrho(\psi(y), F \circ \psi \circ \tau^{-1}(y)) + \\ H(F \circ \psi \circ \tau^{-1}(y), F \circ \alpha \circ \tau^{-1}(y)) &\leq \\ \delta + \lambda d(\psi \circ \tau^{-1}(y), \alpha \circ \tau^{-1}(y)) &\leq \delta + \lambda\varepsilon = \varepsilon.\end{aligned}$$

The right hand side does not depend on y , so $\sup_{y \in Y} \varrho(\psi(y), F \circ \alpha \circ \tau^{-1}(y)) \leq \varepsilon$,

which implies that $\bar{\varrho}(\psi, \Phi(\alpha)) \leq \varepsilon$. Define a set-valued map $\Phi' : X \rightarrow \mathcal{P}_{cl}(X)$ as follows: $\Phi'(\alpha) = \Phi(\alpha) \cap X_{\psi,\varepsilon}$. Since $\Phi'(\alpha) \subset \Phi(\alpha)$ ($\alpha \in X_{\psi,\varepsilon}$), one has

$$\bar{H}(\Phi'(\alpha), \Phi'(\beta)) \leq \bar{H}(\Phi(\alpha), \Phi(\beta)) \leq \lambda \bar{d}(\alpha, \beta),$$

so Φ' is a contracting set-valued mapping. By [8] (see also [7]), this mapping has at least one fixed point $\varphi \in X_{\psi,\varepsilon}$, i.e. $\varphi \in \Phi'(\varphi) \subset \Phi(\varphi)$. The later means that φ is an F -invariant cross-section and $\bar{d}(\varphi, \psi) \leq \varepsilon$. ■

4. Shadowing in contracting relations

Definition 4.1. Given $\varepsilon > 0$, a finite or infinite sequence $\{x_n\} \subset X$ is called an ε -chain for a relation $f : X \rightarrow \mathcal{P}(X)$ (or (ε, f) -chain, or pseudo-chain), if $\varrho(x_{n+1}, f(x_n)) \leq \varepsilon$ for all n .

Remark 4.1. One can easily see that a (ε, f) -chain is a chain for the relation $\bar{V}_\varepsilon \circ f$.

Thus, in addition to providing an abstract framework for generalizing properties of iterates of maps, the study of iteration of relations includes pseudo-chains (ε -chains) as a special case of chains for a neighborhood $\bar{V}_\varepsilon \circ f$ of the relation f .

Definition 4.2. One says that the relation $f : X \rightarrow \mathcal{P}(X)$ has the *Shadowing Property on the subset $A \subset X$* , if, given $\varepsilon > 0$, there exists $\delta > 0$ such that for any δ -chain $\{x_n\} \subset A$ there exists a chain $\{y_n\} \subset X$ such that $d(x_n, y_n) \leq \varepsilon$ for all n (one says that the δ -chain $\{x_n\}$ is ε -shadowed by the chain $\{y_n\}$).

Theorem 4.1. Any contracting relation $f : X \rightarrow \mathcal{P}_{b,cl}(X)$ has the Shadowing Property on X .

Proof. We will reduce the proof of this theorem to Theorem 3.1. For, take $Y = \mathbf{N} = \{1, 2, \dots\}$ and $\tau : \mathbf{N} \rightarrow \mathbf{N}$, $\tau(n) = n + 1$. It is easy to see

that the sequence $\{x_n\} \subset X, n \in \mathbf{N}$, is a chain (respectively, an ε -chain) for f , if and only if the function $\psi : \mathbf{N} \rightarrow \mathbf{N} \times X, \psi(n) = (n, x_n)$, is an invariant (respectively, an ε -invariant) cross-section for the relation $F = (\tau, f)$. Thus, the result follows from Theorem 3.1. ■

Corollary. *If the relation f on X has the Shadowing Property on X , then for any $\varepsilon > 0$ there exists $\delta > 0$ such that for any relation g on X , satisfying $H(f(x), g(x)) < \delta (x \in X)$, any δ -chain $\{x_n\}$ of g is ε -shadowed by some chain $\{y_n\}$ of f .*

Proof. It is known, that, given $\varepsilon > 0$, there exists $\bar{\delta} > 0$ such that any $\bar{\delta}$ -chain $\{x_n\}$ of f is ε -shadowed by some chain $\{y_n\}$ of f .

Take $\delta = \bar{\delta}/2$. Let g be a relation on X , satisfying $H(f(x), g(x)) < \delta (x \in X)$ and $\{x_n\}$ be a δ -chain of g . At the same time, $\{x_n\}$ is a (2δ) -chain of f , since, by Lemma 1.1, we have

$$\varrho(x_{n+1}, f(x_n)) \leq \varrho(x_{n+1}, g(x_n)) + H(g(x_n), f(x_n)) < 2\delta = \bar{\delta}.$$

Hence, $\{x_n\}$, as a $\bar{\delta}$ -chain of f , is ε -shadowed by some chain $\{y_n\}$ of f . ■

Remark 4.2. The Shadowing Property is not a robust property for the relations (even for contractions) in the sense that this property is to be kept under small perturbations of the relation with respect to the metric \tilde{H} . For example, consider $f : \mathbf{R} \rightarrow \mathbf{R}, f(x) = \frac{1}{2}x$, and construct a closed relation g on \mathbf{R} as follows: g is equal to $\text{id}_{\overline{V}}$ on a small closed neighborhood \overline{V} of zero, and to f outside \overline{V} . The relation g has not the Shadowing Property on \mathbf{R} .

Theorem 4.2. *Let $f : X \rightarrow \mathcal{P}_{b,cl}(X)$ be a contracting relation. Then for any $\varepsilon > 0$ there exists $\delta > 0$ such that for any periodic δ -chain $\{x_n\}$ there exists a periodic chain $\{y_n\}$ for f such that $d(x_n, y_n) \leq \varepsilon$ for any $n \in \mathbf{N}$.*

Proof. Fix $\varepsilon > 0$ and take $\delta = (1 - \lambda)\varepsilon$. Let $\{x_n\}$ be a periodic δ -chain for f . Now apply Theorem 3.1 for $Y = \mathbf{Z}_p$ and $\tau : \mathbf{Z}_p \rightarrow \mathbf{Z}_p, \tau(n) \equiv n + 1 \pmod{p}$. Thus, we obtain a p -periodic chain $\{y_n\}$, which ε -shadows $\{x_n\}$. ■

Remark 4.3. It is obvious that every periodic chain for a contracting relation is contained in the attractor. From Theorem 4.2 it follows that any periodic δ -chain must be in a ε -neighborhood of the attractor, since $\varrho(x_n, A) \leq d(x_n, y_n)$ for any $n \in \mathbf{N}$.

Recall that $x \in X$ is named a *periodic point* for the relation $f : X \rightarrow \mathcal{P}(X)$, if there exists a periodic chain, starting at x , or, in other words, if x is a fixed point for the orbit relation $\mathcal{O}f = \bigcup_{n \geq 1} f^n$. Following [4], denote by $|\mathcal{O}f|$ the set of periodic points of f .

Theorem 4.3. For a contracting relation $f : X \rightarrow \mathcal{P}(X)$ the periodic points form a dense subset of the attractor, i.e. $\overline{\mathcal{O}f} = A$.

Proof. For any $\varepsilon > 0$ and any point $x \in A$ we have to find a periodic point $y \in A$ such that $d(x, y) < \varepsilon$. If x is a periodic point it is nothing to prove.

Assume that x is not a periodic point, i.e. $x \notin f^n(x)$ for any $n \in \mathbb{N}$. Because of compactness of A for any $\eta > 0$ there exists n such that $\varrho(x, f^n(x)) \leq \eta$. For $\eta < (1 - \lambda)\varepsilon/2$ take a (finite) chain $x_1 = x, x_2, \dots, x_n \in f^n(x)$ such that $d(x, x_n) \leq \eta$ and extend it up to a periodic η -chain $\{x_1, x_2, \dots, x_n, x_{n+1} = x_1, \dots\}$. In virtue of Theorem 4.2 for this periodic η -chain there exists a periodic chain $\{y_n\}$ such that $d(x_n, y_n) \leq \varepsilon/2 < \varepsilon$ for any $n \in \mathbb{N}$. Thus, $y = y_1$ is the desired periodic point. ■

Theorem 4.4. Let $f : X \rightarrow \mathcal{P}_{b,cl}(X)$ be a contracting relation and A stand for its attractor. Then for any $a \in A$ one has $\overline{\mathcal{O}f(a)} = A$.

Proof. Because of $\mathcal{O}f = \bigcup_{n \geq 1} f^n$, the inclusion $\overline{\mathcal{O}f(a)} \subset A$ follows from the invariance of A and its closedness.

For proving the inverse inclusion we use the Lemma 1.2 and observe that $H(f^n(a), A) = H(f^n(a), f^n[A]) \leq \lambda^n H(a, A) \rightarrow 0$ as $n \rightarrow \infty$.

Thus, for any $b \in A$ and any $\varepsilon > 0$ there exists natural n such that $\varrho(b, f^n(a)) < \varepsilon$.

Due to the compactness of A we have that for any $b \in A$ and any $\varepsilon > 0$ there exist natural n and $a' \in f^n(a)$ such that $d(b, a') < \varepsilon$. This implies the equality $\overline{\mathcal{O}f(a)} = A$. ■

References

- [1] Valero J. *Attractors of parabolic equations without uniqueness*. J. Dynam. Diff. Eq., Vol. 13, 4(2001), 711-744.
- [2] Melnik V.S., Valero J. *On attractors of multivalued semi-flows and differential inclusions*. Set-valued Analysis, 6(1998), 83-111.
- [3] Bobylev N., Emelianov S., Korovin S. *Attractors in control systems*. Different. Uravneniya, 35(1999), 617-622.
- [4] Akin E. *The General Topology of Dynamical Systems*. Amer. Math. Soc., Providence, RI, 1993.
- [5] McGehee R. *Attractors for closed relations on compact Hausdorff spaces*. Indiana U. Math. J., Vol. 41, 4(1992), 1165-1209.
- [6] Barnsley M. *Fractals Everywhere*. Academic Press Professional, Boston, 1988.
- [7] Rus I.A. *Generalized Contractions and Applications*. Cluj University Press, Cluj-Napoca, 2001.
- [8] Nadler S.B. *Multi-valued contraction mappings*. Pacific J. Math., 30(1969), 475-488.

LEVEL SET METHODS FOR A PARAMETER IDENTIFICATION PROBLEM

Bjørn–Ove Heimsund

Department of Mathematics, University of Bergen, Norway

bjorno@math.uib.no

Tony Chan

Department of Mathematics, University of California, Los Angeles

TonyC@college.ucla.edu

Trygve K. Nilssen

Simula Research Laboratory, Oslo, Norway

trygvekn@simula.no

Xue–Cheng Tai

Department of Mathematics, University of Bergen, Norway

tai@math.uib.no

1. Introduction

Consider the partial differential equation

$$\begin{cases} -\nabla \cdot (q(x)\nabla u) &= f & \text{in } \Omega \subset \mathbb{R}^d, \\ u &= 0 & \text{on } \partial\Omega. \end{cases} \quad (1)$$

We want to use observations of the solution u to recover the coefficient $q(x)$. We shall especially treat the case that $q(x)$ has discontinuities and is piecewise constant.

In this work, we shall combine the ideas used in [5] and [2] to use level set methods to estimate the coefficient $q(x)$. The level set method was first proposed in Osher and Sethian [8]. This method associate a two-dimensional closed curve with a two-dimensional function. Extensions to higher dimensions are also easy, see Ambrosio and Soner [1]. The

advantage of using level set method is that it gives a better tool for evolving curves that may disappear, merge with each other, or pinch off with each other.

The level set method has been used for some inverse problems in [4, 10], etc. The work of [5] seems to be the first one to apply the level set idea to estimate the coefficient $q(x)$ from the equation (1). The work of [5] only works when the coefficient $q(x)$ takes two constants values, i.e. they can only handle one level set function. Several approaches have been proposed to use multiple level set functions (not for inverse problems, but for fluid and image problems) [2, 13]. The approach of [2] is easy to implement and has been well tested for image segmentation problems. In this work, we are trying to use the idea of [2] for the parameter estimation problem. See also [11, 12] for some related works in using Heaviside functions to identify shapes and boundaries.

2. Level set methods

Here, we state some of the details of the level set idea. Let Γ be a closed curve in Ω . Associated with Γ , we define ϕ as a signed distance function by

$$\phi(x) = \begin{cases} \text{distance}(x, \Gamma), & x \in \text{interior of } \Gamma \\ -\text{distance}(x, \Gamma), & x \in \text{exterior of } \Gamma, \end{cases}$$

In many applications, the movement of the curve Γ can be described by a partial differential equation of the function ϕ . The function ϕ is called a level set function for Γ . In fact, ϕ is the unique viscosity solution null on Γ for the following partial differential equation

$$|\nabla \phi| = 1, \quad \text{in } \Omega. \tag{2}$$

In this work, we shall use the level set method to identify the coefficient q which is assumed to be piecewise constant. First look at a simple case, i.e. assume that q has a constant value q_1 inside a closed curve Γ and is another constant q_2 outside the curve Γ . Utilizing the Heaviside function $H(\phi)$, which is equal to 1 for positive ϕ and 0 elsewhere, it is easy to see that q can be represented as

$$q = q_1 H(\phi) + q_2 (1 - H(\phi)). \tag{3}$$

In order to identify the coefficient q , we just need to identify a level set function ϕ and the piecewise constant values q_i .

If a function has many pieces, then we need to use multiple level set functions. This idea was introduced in Chan and Vese [2]. Assume that we have two closed curves Γ_1 and Γ_2 , and we associate the

two level set functions $\phi_j, j = 1, 2$ with these curves. Then the domain Ω is divided into the four pars $\Omega_{++} = \{x \in \Omega, \phi_1 > 0, \phi_2 > 0\}$, $\Omega_{+-} = \{x \in \Omega, \phi_1 > 0, \phi_2 \leq 0\}$, $\Omega_{-+} = \{x \in \Omega, \phi_1 \leq 0, \phi_2 > 0\}$, $\Omega_{--} = \{x \in \Omega, \phi_1 \leq 0, \phi_2 \leq 0\}$.

Allowing some of the subdomains defined above to be empty, we can easily handle the case that the zero level set curves could merge, split or disappear. Using the Heaviside function again, we can express q with possibly up to four pieces with constant values as

$$\begin{aligned} q = & q_1 H(\phi_1) H(\phi_2) + q_2 H(\phi_1)(1 - H(\phi_2)) + \\ & + q_3(1 - H(\phi_1)) H(\phi_2) + q_4(1 - H(\phi_1))(1 - H(\phi_2)). \end{aligned} \quad (4)$$

By generalizing, we see that n level set functions give the possibility of 2^n regions. In that case, q would look like

$$\begin{aligned} q = & q_1 H(\phi_1) H(\phi_2) \cdots H(\phi_n) + \\ & + q_2(1 - H(\phi_1)) H(\phi_2) \cdots H(\phi_n) + \\ & \vdots \\ & + q_{2^n}(1 - H(\phi_1))(1 - H(\phi_2)) \cdots (1 - H(\phi_n)). \end{aligned} \quad (5)$$

Even if we need less than 2^n distinct regions, we can still use n level set functions since some subdomains may be empty. In using such a representation, we need to determine the maximum number of level set functions we want to use before we start.

For many practical applications, such kind of a priori information is often available or is chosen according the measurements that are available to us. Also, to ensure ellipticity of equation (1), we need each q_i to be positive, that is, we assume that there exist $0 < a_i < b_i < \infty$ that are known a priori such that $q_i \in [a_i, b_i]$.

3. The parameter identification problem

We shall try to identify the coefficient q from a measurement of u on a subdomain $\hat{\Omega}$. We shall perform the numerical tests both for the case that $\hat{\Omega} = \Omega$ and the case that $\hat{\Omega} \subset \Omega$. In case that $\hat{\Omega} = \Omega$, existence and uniqueness of the inverse problem is already known. For general cases, studies about existence and uniqueness are still missing in the literature. In this work, the parameter identification problem is formulated as a least-square minimization problem and then we propose to use the augmented Lagrangian method to solve the least-squares minimization problem with the equation as constraint.

We start by defining the equation error of equation (1) as $e = e(q, u) \in H_0^1(\Omega)$ which is the variational solution of

$$(\nabla e, \nabla v) = (q \nabla u, \nabla v) - (f, v), \quad \forall v \in H_0^1(\Omega). \quad (6)$$

Here and later (\cdot, \cdot) denotes the L^2 -innerproduct over Ω . We will also use the notation $\|\cdot\|$ to denote the associated norm. For a given q and a given u , we say that they satisfy the equation (1) if and only if $e(q, u) = 0$. In order to solve our inverse problem, we shall try to find a q and u such that $e(q, u) = 0$ and also fits the measurements \hat{u} best among all admissible functions q and u . Using the level set functions, the coefficient q will be represented as functions of the level set functions ϕ_j and the piecewise values q_i and the minimization problem we need to solve takes the form

$$\min_{q_i, \phi_j, u} \left(\frac{1}{2} \|u - \hat{u}\|_{L^2(\hat{\Omega})}^2 + \beta \sum_{i=1}^n \int_{\Omega} |\nabla H(\phi_j)| dx \right), \quad (7)$$

under the conditions $e(q, u) = 0, \quad |\nabla \phi_j| = 1, \forall j.$

In the above q is a function of ϕ_j and q_i . The constraint $e(q, u) = 0$ makes sure the equation error is zero. The first term tries to minimize the deviation between the calculated u and measured \hat{u} , while $\sum_{i=1}^n \int_{\Omega} |\nabla H(\phi_j)| dx$ in the second term is referred to as a regularization term. In case that Ω is a one-dimensional domain, then $\int_{\Omega} |\nabla H(\phi_j)| dx$ equals to the number of points that the level set functions ϕ_j equals zero. If Ω is two-dimensional, it is the length of the zero level set curves of ϕ_j . For three-dimensional cases, then it is the area of the zero level set surfaces of ϕ_j .

To solve (7), we use the augmented Lagrangian formulation, and the corresponding Lagrangian functional $L : R^{2^n} \times [Lip(\Omega)]^n \times H_0^1(\Omega) \times H_0^1(\Omega) \mapsto R$ is

$$\begin{aligned} L(q_i, \phi_j, u, \lambda) = & \frac{1}{2} \|u - \hat{u}\|_{L^2(\hat{\Omega})}^2 + \\ & + \beta \sum_{j=1}^n \int_{\Omega} |\nabla H(\phi_j)| dx + \frac{c}{2} \|\nabla e\|^2 - (\nabla \lambda, \nabla e). \end{aligned} \quad (8)$$

The Lagrangian multiplier λ is only trying to enforce the equation constraint $e(q, u) = 0$. The other constraints $|\nabla \phi_j| = 1$ will be enforced by some other methods well developed for the level set methods. Due to the fact that the ϕ_j 's are the viscosity solutions for the Eikonal equation, it is not easy to enforce them by the Lagrangian multiplier method.

In order to find a minimizer for the minimization problem (7), we shall use an algorithm of the type of the Lancelot method which will be given later in this section. The algorithm needs the derivatives of the Lagrangian functional with respect to the minimization variables.

3.1. Calculation of $\nabla_{q_i} L$

The derivative of L with respect to q_i is

$$\frac{\partial L}{\partial q_i} = c \left(\nabla e, \nabla \frac{\partial e}{\partial q_i} \right) - \left(\nabla \lambda, \nabla \frac{\partial e}{\partial q_i} \right) = \left(\nabla \frac{\partial e}{\partial q_i}, \nabla (ce - \lambda) \right).$$

From (6), it follows that the derivative of ∇e with respect to q_i is

$$\left(\nabla \frac{\partial e}{\partial q_i}, \nabla v \right) = \left(\frac{\partial q}{\partial q_i} \nabla u, \nabla v \right).$$

Taking v to be $ce - \lambda$ gives

$$\frac{\partial L}{\partial q_i} = \left(\frac{\partial q}{\partial q_i} \nabla u, \nabla (ce - \lambda) \right).$$

It is trivial to calculate the derivative of q with respect to q_i when using equation (5).

3.2. Calculation of $\nabla_{\phi_j} L$

For clarity of the presentation, we shall first calculate the Gateaux derivative of the regularization term, i.e. we first calculate the Gateaux derivative for the following functional

$$R(\phi_j) = \int_{\Omega} |\nabla H(\phi_j)| dx = \int_{\Omega} \delta(\phi_j) |\nabla \phi_j| dx.$$

Here and later, δ denotes the Dirac-function. To get the derivative of R with respect to ϕ_j in the direction μ_j , we proceed

$$\frac{\partial R}{\partial \phi_j} \cdot \mu_j = \int_{\Omega} \delta'(\phi_j) \mu_j |\nabla \phi_j| dx + \int_{\Omega} \delta(\phi_j) \frac{\nabla \phi_j}{|\nabla \phi_j|} \cdot \nabla \mu_j dx.$$

Applying Greens formula to the last term which can be theoretically verified by replacing the delta function by a smooth function and then passing to the limit, we will get that

$$\begin{aligned} \frac{\partial R}{\partial \phi_j} \cdot \mu_j &= \int_{\Omega} \delta'(\phi_j) \mu_j |\nabla \phi_j| dx - \int_{\Omega} \nabla \cdot \left(\delta(\phi_j) \frac{\nabla \phi_j}{|\nabla \phi_j|} \right) \mu_j dx \\ &= \int_{\Omega} \delta'(\phi_j) \mu_j |\nabla \phi_j| dx - \int_{\Omega} \left(\delta'(\phi_j) \frac{|\nabla \phi_j|^2}{|\nabla \phi_j|} \mu_j + \delta(\phi_j) \mu_j \nabla \cdot \frac{\nabla \phi_j}{|\nabla \phi_j|} \right) dx \quad (9) \\ &= - \int_{\Omega} \delta(\phi_j) \mu_j \nabla \cdot \frac{\nabla \phi_j}{|\nabla \phi_j|} dx, \end{aligned}$$

which indicates that

$$\frac{\partial R}{\partial \phi_j} = -\delta(\phi_j) \nabla \cdot \frac{\nabla \phi_j}{|\nabla \phi_j|}.$$

Denote the Gateaux derivative of L with respect to ϕ_j in the direction μ_j as $\frac{\partial L}{\partial \phi_j} \cdot \mu_j$. The Gateaux derivative in this case is

$$\begin{aligned} \frac{\partial L}{\partial \phi_j} \cdot \mu_j &= c \left(\nabla e, \nabla \left(\frac{\partial e}{\partial \phi_j} \cdot \mu_j \right) \right) - \left(\nabla \lambda, \nabla \left(\frac{\partial e}{\partial \phi_j} \cdot \mu_j \right) \right) + \beta \frac{\partial R}{\partial \phi_j} \cdot \mu_j \\ &= \left(\nabla \frac{\partial e}{\partial \phi_j} \cdot \mu_j, c \nabla e - \nabla \lambda \right) + \beta \frac{\partial R}{\partial \phi_j} \cdot \mu_j. \end{aligned}$$

The derivative of e with respect to ϕ_j in the direction μ_j is

$$\left(\nabla \left(\frac{\partial e}{\partial \phi_j} \cdot \mu_j \right), \nabla v \right) = \left(\frac{\partial q}{\partial \phi_j} \cdot \mu_j \nabla u, \nabla v \right),$$

Taking v to be $ce - \lambda$, we get that

$$\frac{\partial L}{\partial \phi_j} \cdot \mu_j = \left(\frac{\partial q}{\partial \phi_j} \cdot \mu_j, \nabla u \cdot \nabla (ce - \lambda) \right) + \beta \frac{\partial R}{\partial \phi_j} \cdot \mu_j. \quad (10)$$

From (5), it is easy to calculate the Gateaux derivative $\frac{\partial q}{\partial \phi_j} \cdot \mu_j$. For simplicity of the presentation, let us take the case that we only have two level set functions. Then q takes the form (4). Consequently, the Gateaux derivative for the function ϕ_j in a direction μ_j is

$$\frac{\partial q}{\partial \phi_1} \cdot \mu_1 = [(q_1 - q_3)H(\phi_2) + (q_2 - q_4)(1 - H(\phi_2))] \delta(\phi_1) \mu_1 \quad (11)$$

$$\frac{\partial q}{\partial \phi_2} \cdot \mu_2 = [(q_1 - q_2)H(\phi_1) + (q_3 - q_4)(1 - H(\phi_1))] \delta(\phi_2) \mu_2. \quad (12)$$

3.3. Calculation of $\nabla_u L$

We perturb u to $u + \epsilon w$ and try to calculate the Gateaux derivative of L with u in the direction w . First note that

$$\left(\nabla \left(\frac{\partial e}{\partial u} \cdot w \right), \nabla v \right) = (q \nabla w, \nabla v), \quad \forall v \in H_0^1(\Omega),$$

and

$$\begin{aligned} \frac{\partial L}{\partial u} \cdot w &= (u - \hat{u}, w)_{L^2(\hat{\Omega})} + c \left(\nabla e, \nabla \left(\frac{\partial e}{\partial u} \cdot w \right) \right) - \left(\nabla \lambda, \nabla \left(\frac{\partial e}{\partial u} \cdot w \right) \right) \\ &= (u - \hat{u}, w)_{L^2(\hat{\Omega})} + \left(\nabla (ce - \lambda), \nabla \left(\frac{\partial e}{\partial u} \cdot w \right) \right). \end{aligned}$$

Combining the above two equalities, it is true that

$$\frac{\partial L}{\partial u} \cdot w = (u - \hat{u}, w)_{L^2(\hat{\Omega})} + (\nabla(ce - \lambda), q\nabla w).$$

This indicates that

$$\frac{\partial L}{\partial u} = (u - \hat{u})\chi_{\hat{\Omega}} - \nabla \cdot (q\nabla(ce - \lambda)),$$

where $\chi_{\hat{\Omega}}$ is the characteristic function for the subdomain $\hat{\Omega}$, i.e. $\chi_{\hat{\Omega}}(x) = 1$ if $x \in \hat{\Omega}$ and $\chi_{\hat{\Omega}}(x) = 0$ if $x \notin \hat{\Omega}$.

3.4. An algorithm of Lancelot type

To solve the minimization problem (8) we will use a Lancelot type algorithm, as described in Conn, Gould and Toint [3]. In our case, the algorithm can be written as follows

Algorithm 1 Choose q_i^0 , ϕ_j^0 , u^0 as initial guess for the solution, and set $\lambda^0 = 0$, $k = 0$. Also chose initial tolerances $\epsilon_m > 0$, $\epsilon_e > 0$, and the parameters $c > 0$, $\beta > 0$, $\omega > 1$.

Then iteratively do the following steps:

1. Find an approximative minimum $(q_i^{k+1}, \phi_j^{k+1}, u^{k+1})$ of equation (8) such that $\|\nabla_{q_i^{k+1}, \phi_j^{k+1}, u^{k+1}} L\| \leq \epsilon_m$.
2. If $\|\nabla e\| < \epsilon_e$, update λ by $\lambda^{k+1} = \lambda^k - ce(q_i^{k+1}, \phi_j^{k+1}, u^{k+1})$, else update c by $c \leftarrow \omega c$.
3. Decrease ϵ_m , ϵ_e , and set $k \leftarrow k + 1$.

In the first step of the iteration, we actually try to solve the minimization problem. Since λ is not solved for, neither do we need to drive $\|\nabla L\|$ to zero. The second step checks to see if the equation error is sufficiently small, and if so is the case, λ may be updated. Should the error not be small, the penalty parameter c is increased, making the augmented Lagrangian functional more dominated by the $\|\nabla e\|^2$ term. With these steps, we can decrease the tolerances, and go to the next iteration.

For the minimization problem, one may chose any suitable method for nonlinear, unconstrained problems, such as the method of steepest descent, the non-linear conjugate gradients, or the Quasi-Newton methods. Steepest descent methods, while being simple to implement, are known for giving low performance, and the Quasi-Newton methods needs careful implementation and provisions to limit memory usage, but they often

yield excellent performance. In our case, the nonlinear conjugate gradients method was found suitable. It has good performance, low memory usage, and is easy to implement. See [7] for more information regarding these methods.

There are some aspects of our minimization that require additional explanation. We compute the composite derivative of L , that is a vector consisting of $\nabla_{q_i^{k,l}}$, then $\nabla_{\phi_j^{k,l}}$ and finally $\nabla_{u^{k,l}}$. This makes us minimize L with all the unknowns simultaneously. The reason for doing this instead of optimizing each unknown separately is that the latter approach is equivalent to a coordinate descent method, giving poor convergence. However, there may be problems finding an optimal steplength for this composite search direction, ie. ϕ_j may require a smaller steplength than u . By defining

$$\tilde{u} = su, \quad s > 0,$$

and using \tilde{u} in the optimization, the rate of convergence may be improved. We have in our experiments chosen s experimentally.

We also need to enforce constraints on q_i and ϕ_j . The former is just to set $q_i = \min\{\max\{q_i, a_i\}, b_i\}$, while for the latter we would try to ensure that

$$|\nabla \phi_j^{k+1}| = 1, \quad \phi_j^{k+1} = 0 \text{ on } \Gamma_j^k.$$

For a one-dimensional problem, this can easily be done, but in higher dimensions we can use methods described in Osher and Fedkiw [9], and Smereka, Sussman, Fatemi and Osher [6]. There are fast and cheap algorithms to solve this problem, see [6, 9].

4. Numerical experiments

In our numerical tests, we will consider three cases. First, the q_i 's are all known, and we try to identify ϕ_j and u ; then we perturb q_i , and attempt to identify ϕ_j and u with q_i fixed. Thirdly, we add noise to \hat{u} , and try to identify ϕ_j and u while q_i is known.

The equation we will use is

$$\begin{aligned} -\nabla \cdot (q \nabla u) &= 2\pi^2 \sin(\pi x) \sin(\pi y), & \text{in } \Omega \\ u &= 0, & \text{on } \partial\Omega. \end{aligned} \tag{13}$$

Here, $\Omega = (0, 1) \times (0, 1)$. All experiments are done with a uniform, 2D mesh of Ω , with 24×24 elements. The numerical parameters used are $\beta = 10^{-7}$, $c = 5 \cdot 10^{-6}$, $\omega = 1.1$, $s = 25$. All the Figures show the zero level sets of the exact and of the computed ϕ_1 (in the lower-left corner) and ϕ_2 (in the upper-right corner) in the various situations that we have considered.

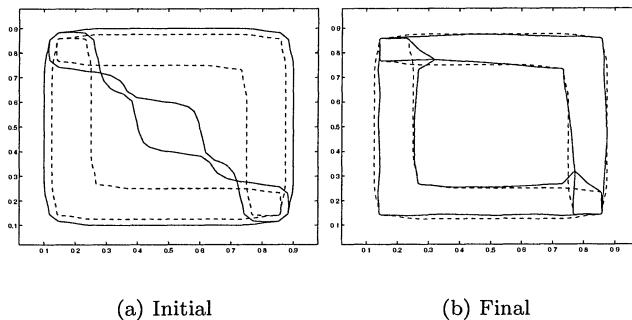


Figure 1. Drawing of the zero level set curves for equation (13). The dashed lines are the exact zero level sets, and the solid lines are the computed solution. This is the first case, and convergence was attained after about 300 iterations. Also, $\|u_h - u\|_2 \approx 1.1906 \cdot 10^{-4}$ at time of convergence.

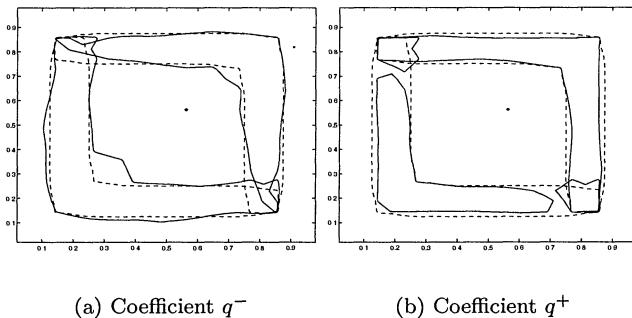


Figure 2. The dashed lines are the exact zero level sets, and the solid lines are the computed solution. This is the second case, with perturbed q_i 's and we start with the same initial levelsets as in the first case. For q^- , it took about 350 iterations to converge, and $\|u_h - u\|_2 \approx 1.3027 \cdot 10^{-4}$ at that time, while for q^+ it only took about 100 iterations, and here $\|u_h - u\|_2 \approx 1.1182 \cdot 10^{-4}$. Since the discontinuities are smaller in this latter case, this quicker convergence is expected.

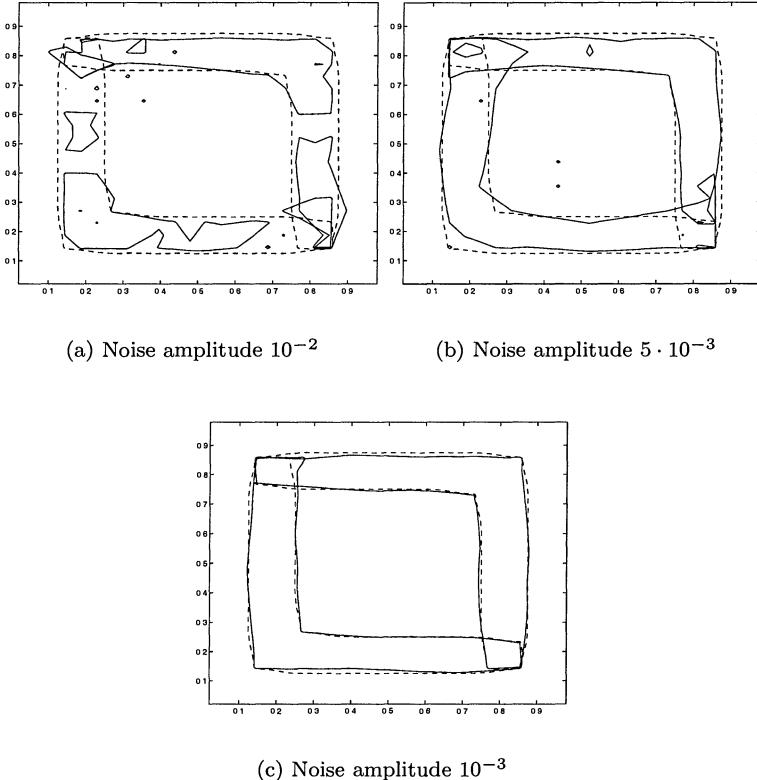


Figure 3. The dashed lines are the exact zero level sets, and the solid lines are the computed solution. Here is the third case, with noise added to \hat{u} . After about 100 iterations, we had convergence, and the error $\|u_h - u\|_2$ was about $1.7251 \cdot 10^{-4}$, $1.5019 \cdot 10^{-4}$ and $1.4223 \cdot 10^{-4}$ for noise amplitudes of 10^{-2} , $5 \cdot 10^{-3}$ and 10^{-3} respectively. More noise made it generally impossible to get convergence, and with less noise the solution was not distinguishable from the case without noise.

We shall use q given as follows. $q_1 = 1$ when $\phi_1, \phi_2 \leq 0$, $q_2 = 1 + 2^{1/3}$ when $\phi_1 \leq 0, \phi_2 > 0$, $q_3 = 1 + 2^{2/3}$ when $\phi_1 > 0, \phi_2 \leq 0$, and finally $q_4 = 3$ when $\phi_1, \phi_2 > 0$. ϕ_1 is positive within the union of the following rectangles

$$\{x, y : 3/4 < x < 7/8, 1/8 < y < 7/8\} \cup \{x, y : 1/8 < x < 7/8, 3/4 < y < 7/8\},$$

and ϕ_2 is likewise positive within this union

$$\{x, y : 1/8 < x < 1/4, 1/8 < y < 7/8\} \cup \{x, y : 1/8 < x < 7/8, 1/8 < y < 1/4\}.$$

In all cases, we let $\hat{\Omega}$ extend from the boundary of Ω and into the interior by a length of $1/3$ from all sides. In the remainder, we let $\hat{\Omega}$ be coarser than Ω by a factor of two. Thus we have complete observations \hat{u} along the boundary, and coarser observations in the interior. Note that $\|\nabla u\| \approx 0$ near the center of Ω . Because of this, we cannot easily identify q in the center since for $\|\nabla u\| = 0$, q is no longer unique.

Solving for the first case yields the results in Figure 1, which are quite accurate.

For the second case, we will use the two sets of q_i coefficients q^- and q^+ , and they are given as follows. $q_1^\pm = q_1 \pm 0.1$, $q_2^\pm = q_2 \pm 0.051$, $q_3^\pm = q_3 \mp 0.05$, $q_4^\mp = q_4 \pm 0.1$.

Note that the perturbations in q^- create larger jumps, while in q^+ the jumps are smaller. The results are in Figure 2, and we see that it is easier to identify q^+ rather than q^- due to smaller jumps. Also note that the deviations from Figure 1 are small.

We now come to the third case where we will add noise to our observations, and try find ϕ_j and u . Adding normally distributed noise of varying magnitude gives us the results in Figure 3. Having larger noise-magnitude than 10^{-2} makes it generally hard to get convergence, while a noise-magnitude of less than 10^{-3} hardly makes much of an impact on our case.

References

- [1] L. Ambrosio and H. M. Soner. Level set approach to mean curvature flow in arbitrary codimension. *J. Diff. Geom.*, 43(4):693–737, 1996.
- [2] Tony F. Chan and Luminita A. Vese. A new multiphase level set framework for image segmentation via the Mumford and Shah model. Technical report, CAM Report 01-25, UCLA, April 2001.
- [3] A. R. Conn, N. I. M. Gould, and P. L. Toint. *LANCELOT, a FORTRAN package for Large-scale nonlinear optimization (Release A)*. no. 17 in Springer series in Computational mathematics. Springer-Verlag, New-York, 1992.
- [4] D. C. Dobson and F. Santosa. An image enhancement technique for electrical impedance tomography. *Inverse problems*, 10:317–334, 1994.

- [5] K.Ito, K. Kunisch, and Z. Li. Level-set function approach to an inverse interface problem. *Inverse Problems*, 17:1225–1242, 2001.
- [6] P. Smereka M. Sussman, E. Fatemi and S. Osher. An improved level set method for incompressible two-phase flow. *Computers and Fluids*, 27:663–680, 1998.
- [7] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer-Verlag, 1999.
- [8] S. Osher and J. A. Sethian. Fronts propagating with curvature dependent speed: Algorithms based on Hamilton-Jacobi formulations. *J. Comput. Phys.*, 79:12–49, 1988.
- [9] Stanley Osher and Ronald R. Fedkiw. Level set methods. Technical report, CAM Report 00-08, UCLA, February 2000.
- [10] Fadil Santosa. A level-set approach for inverse problems involving obstacles. *ESAIM Contrôle Optim. Calc. Var.*, 1:17–33 (electronic), 1995/96.
- [11] W.B.Liu, P.Neittaanmaki, and D.Tiba. Sur les problemes d'optimisation structurelle. *CRAS, Ser.I Math.*, 331:101–106, 2000.
- [12] W.B.Liu, P.Neittaanmaki, and D.Tiba. Existence for shape optimization problems in arbitrary dimension. *SIAM J. Control and Optimiz.*, to appear, 2002.
- [13] Hong-Kai Zhao, T. Chan, B. Merriman, and S. Osher. A variational level set approach to multiphase motion. *J. Comput. Phys.*, 127(1):179–195, 1996.

FACTORIZATION OF ELLIPTIC BOUNDARY VALUE PROBLEMS: THE *QR* APPROACH

Jacques Henry

INRIA

BP 105, 78153 Le Chesnay FRANCE

Jacques.Henry@inria.fr

1. Introduction

In this paper we describe and develop a method first proposed by Angel and Bellman ([1]) to factorize a second order elliptic boundary value problem in the product of two first order decoupled initial value problems by invariant embedding. For the sake of simplicity we consider a domain Ω of \mathbb{R}^n which is a cylinder $]0, 1[\times \mathcal{O}$ and the Laplacian Δ as elliptic operator. We denote x the coordinate along the first axis which is also the axis of the cylinder and y the $n - 1$ other coordinates. The section $\mathcal{O} \subset \mathbb{R}^{n-1}$ is bounded and has a smooth boundary. We denote $\Sigma =]0, 1[\times \partial\mathcal{O}$ the lateral boundary of the cylinder and $\Gamma_0 = \{0\} \times \mathcal{O}$, $\Gamma_1 = \{1\} \times \mathcal{O}$ the two faces of the cylinder. We consider the problem

$$(\mathcal{P}_0) \begin{cases} -\Delta u = f & \text{in } \Omega, \\ u|_{\Sigma} = 0, \quad -\frac{\partial u}{\partial x}|_{\Gamma_0} = u_0, \quad u|_{\Gamma_1} = u_1. \end{cases}$$

The case of a Dirichlet boundary condition on Γ_0 will also be considered in section 5. The problem is embedded in a family of similar problems in the subcylinders $]0, s[\times \mathcal{O}$. Let $Q(s)$ be the Dirichlet-to-Neumann map on the section $x = s$. We prove that the boundary value problem for the Poisson equation can be factorized as:

$$-\left(\frac{d}{dx} + Q\right)\left(\frac{d}{dx} - Q\right)u = f$$

each of the first order problem having an initial value given at $x = 0$ or $x = 1$. Furthermore the operator Q satisfies the Riccati equation

$$\frac{dQ}{dx} - Q^2 = \Delta_y, \quad Q(0) = 0.$$

where Δ_y is the Laplacian on the section \mathcal{O} . A control problem equivalent to the Poisson problem whose time variable is the x -coordinate is presented. The previous Riccati equation yields the optimal feedback for this control problem.

The previous factorization of the Poisson problem can be viewed as an infinite dimensional extension of the Gauss LU block factorization. We also present a similar extension of the QR factorization. It first uses a factorization of the normal equation from which the orthogonal operator can be derived. The triangular part takes the form of a second order in x initial value problem.

Finally we present an optimal control problem with an elliptic state equation. We show that the factorization and uncoupling of both the state and adjoint state can be achieved together.

2. Factorization of the state equation

We briefly recall the factorization of the state equation from [2]. Using the technique of invariant embedding introduced by R. Bellman (see [1]), we embed problem (\mathcal{P}_0) in a family of similar problems $(\mathcal{P}_{s,h})$ defined on $\Omega_s =]0, s[\times \mathcal{O}$ for $s \in]0, 1]$. For each problem we impose the Dirichlet boundary condition $u|_{\Gamma_s} = h$, where $\Gamma_s = \{s\} \times \mathcal{O}$.

$$(\mathcal{P}_{s,h}) \left\{ \begin{array}{l} -\Delta u = f \quad \text{in } \Omega_s, \\ u|_{\Sigma} = 0, \quad -\frac{\partial u}{\partial x}|_{\Gamma_0} = u_0, \quad u|_{\Gamma_s} = h. \end{array} \right.$$

For every $s \in]0, 1]$ we define the Dirichlet-to-Neumann (DtN) map $Q(s)$ by $Q(s)h = \frac{\partial u}{\partial x}|_{\Gamma_s}$, with f and u_0 set to zero. By linearity of $(\mathcal{P}_{s,h})$ we have $\frac{\partial u}{\partial x}|_{\Gamma_s} = Q(s)h + w(s)$.

Furthermore, the solution u of (\mathcal{P}_0) restricted to $]0, s[$ satisfies $(\mathcal{P}_{s,u|_{\Gamma_s}})$ for $s \in]0, 1[$ so that

$$\frac{\partial u}{\partial x}(x, y) = (Q(x)u|_{\Gamma_x})(y) + (w(x))(y). \quad (1)$$

Then, by formally taking the derivative with respect to x of this formula, we obtain $\frac{\partial^2 u}{\partial x^2} = -\Delta_y u - f = \frac{dQ}{dx}u + Q\frac{\partial u}{\partial x} + \frac{\partial w}{\partial x}$, where Δ_y is the $(n-1)$ -dimensional Laplacian on \mathcal{O} . Therefore substituting $\frac{\partial u}{\partial x}$ from equation (1)

$$0 = \left(\frac{dQ}{dx} + Q^2 + \Delta_y \right) u + \frac{\partial w}{\partial x} + Qw + f,$$

and then, since u is arbitrary, we obtain the decoupled system

$$\begin{cases} \frac{dQ}{dx} + Q^2 + \Delta_y = 0, & Q(0) = 0, \\ \frac{dw}{dx} + Qw = -f, & w(0) = -u_0, \\ -\frac{du}{dx} + Qu = -w, & u(1) = u_1. \end{cases} \quad (2)$$

The initial conditions for Q and w at $x = 0$ are obtained from the boundary conditions for u at Γ_0 and from (1) and similarly for the initial conditions for u at $x = 1$. Let us stress that Q is an operator on functions in y depending on x which satisfies a Riccati equation. The system (2) is decoupled because one can integrate the first two equations in x from 0 to 1 giving Q and w , then u is obtained by the integration backwards of the third equation. Formally, we have factorized $-\Delta u = f$ as

$$-\left(\frac{d}{dx} + Q\right)\left(\frac{d}{dx} - Q\right)u = f. \quad (3)$$

Since Q is self adjoint (see [2]), it is clear that the two factors are adjoint of each other. Also, as Q is coercive, the equations for w and u are of parabolic type. In the particular case of the Poisson equation in a cylinder it can be shown that Q and Δ_y commute.

3. Properties of Q

The precise properties of the DtN map Q and the meaning of the Riccati equation (2) are studied in [2] as continuous operator and in [3] in a Hilbert-Schmidt framework. Here we just briefly recall the functional framework used and the main results. We denote X

$$X = \{u \in H^1(\Omega) \mid u|_{\Sigma} = 0\} \equiv L^2(0, 1; H_0^1(\mathcal{O})) \cap H^1(0, 1; L^2(\mathcal{O})).$$

From [5], we introduce the 1/2 interpolate between $L^2(\mathcal{O})$ and $H_0^1(\mathcal{O})$

$$H_{00}^{1/2}(\mathcal{O}) = [L^2(\mathcal{O}), H_0^1(\mathcal{O})]_{1/2}.$$

Then from [5], we have $X \subset C([0, 1], H_{00}^{1/2}(\mathcal{O}))$, which allows to define the trace of $u \in X$, on Γ_s , $u|_s \in H_{00}^{1/2}(\mathcal{O})$. Assuming for the sake of simplicity $u_1 = 0$ (otherwise f and u are translated), we defin $X_0 = \{u \in X \mid u|_1 = 0\}$, and the variational formulation of (\mathcal{P}_0) reads

$$\int_{\Omega} \nabla u \cdot \nabla \varphi \, dx \, dy = \int_{\Omega} f \varphi \, dx \, dy + \langle u_0, \varphi|_0 \rangle_{H_{00}^{1/2}(\mathcal{O})' \times H_{00}^{1/2}(\mathcal{O})}$$

for all $\varphi \in X_0$. Then $Q \in L^{\infty}(0, 1; \mathcal{L}(H_0^1(\mathcal{O}), L^2(\mathcal{O})))$ satisfies the Riccati equation(2) in the following sense

$$\frac{d}{dx}(Q(x)h, \bar{h}) + (Q(x)h, Q(x)\bar{h}) = (\nabla_y h, \nabla_y \bar{h}) \text{ in } \mathcal{D}'([0, 1]), \forall h, \bar{h} \in H_0^1(\mathcal{O}).$$

We also get the properties a.e.

$Q(x) \in \mathcal{L}(H^{3/2}(\mathcal{O}) \cap H_{00}^{1/2}(\mathcal{O}), H_{00}^{1/2}(\mathcal{O}))$ and

$Q \in \mathcal{L}(H_{00}^{1/2}(\mathcal{O}), H_{00}^{1/2}(\mathcal{O})') \cap \mathcal{L}(L^2(\mathcal{O}), H^{-1}(\mathcal{O}))$; $Q(x)$ is self-adjoint and coercive on $H_{00}^{1/2}(\mathcal{O})$ for $x > 0$. Then w and u are defined in X_0 by

$$\begin{aligned} \left(\frac{dw}{dx}, h \right) + (Qw, h) &= -(f, h), \quad \forall h \in L^2(\mathcal{O}), \quad w(0) = -u_0, \\ \left(-\frac{du}{dx}, h \right) + (Qu, h) &= -(w, h), \quad \forall h \in L^2(\mathcal{O}), \quad u(a) = 0. \end{aligned}$$

Hence $\frac{dw}{dx} \in L^2(0, 1; L^2(\mathcal{O}))$, $\frac{du}{dx} \in L^2(0, 1; L^2(\mathcal{O}))$. These are variational formulation of parabolic type problems.

4. Optimal control problem associated to the boundary value problem.

In this section we show the relation with Riccati equations appearing in optimal control theory (see for instance [4]). In fact we show that problem (P_0) can be formulated as an optimal control problem. We use the operator Q and the function w defined in Section 2, with $u_0 = 0$ (for the sake of simplicity). Let us consider the control space $\mathcal{U} = L^2(\Omega)$. For every $v \in \mathcal{U}$ the state $u(v) \in H^1(0, 1; L^2(\mathcal{O}))$ is solution of

$$\begin{cases} \frac{\partial u}{\partial x} = v & \text{in } \Omega, \\ u(1) = u_1. \end{cases} \quad (4)$$

We also denote $\mathcal{U}_{ad} = \{v \in \mathcal{U} : u(v) \in X_{u_1}\}$ the space of admissible controls, where $X_{u_1} = \{h \in L^2(0, 1; H_0^1(\mathcal{O})) \cap H^1(0, 1; L^2(\mathcal{O})) : h(1) = u_1\}$. The desired state u_d is given almost everywhere in x by the solution of the family of (n-1) dimensional problems

$$\begin{cases} -\Delta_y u_d(x) = f(x) & \text{in } \mathcal{O} \\ u_d|_{\partial\mathcal{O}} = 0. \end{cases} \quad (5)$$

Then u_d belongs to $L^2(0, 1; H_0^1(\mathcal{O}))$. Now we look for $u \in \mathcal{U}_{ad}$ such that $J(u) = \inf_{v \in \mathcal{U}_{ad}} J(v)$, where, for every $v \in \mathcal{U}_{ad}$,

$$J(v) = \int_0^1 \|\nabla_y u(v) - \nabla_y u_d\|_{L^2(\mathcal{O})}^2 dx + \int_0^1 \int_{\mathcal{O}} v^2 dx dy. \quad (6)$$

At this point we have the problem that \mathcal{U}_{ad} is not a closed subset of $L^2(\Omega)$ and therefore we cannot use directly the classic techniques (see, for instance, [4]) in order to solve this problem. Nevertheless, since

$\mathcal{U}_{ad} = \{\frac{\partial h}{\partial x} : h \in X_{u_1}\}$, $J(u) = \inf_{v \in \mathcal{U}_{ad}} J(v) = \inf_{h \in X_{u_1}} \bar{J}(h) = \bar{J}(u)$,
where $\frac{\partial h}{\partial x} = u$ and

$$\bar{J}(h) = \int_0^1 \|\nabla_y h - \nabla_y u_d\|_{L^2(\mathcal{O})}^2 dx + \int_0^1 \int_{\mathcal{O}} \left| \frac{\partial h}{\partial x} \right|^2 dx dy. \quad (7)$$

Now, X_{u_1} is a closed convex set in the Hilbert space $L^2(0, 1; H_0^1(\mathcal{O})) \cap H^1(0, 1; L^2(\mathcal{O}))$ and $\bar{J}(h)^{1/2}$ is a norm of that space. Then (see Theorem 1.3 of chapter I of [4]) there exists a unique $u \in X_{u_1}$ satisfying $\bar{J}(u) = \inf_{h \in X_{u_1}} \bar{J}(h)$, which is uniquely determined by

$$\bar{J}'(u)(h) = 0 \quad \forall h \in X_0. \quad (8)$$

Let us show that u is solution of (\mathcal{P}_0) . Developping (7), one gets

$$\bar{J}(u) = \int_{\Omega} |\nabla u|^2 dx - 2 \int_{\Omega} \nabla_y u \cdot \nabla_y u_d dx + \int_{\Omega} |\nabla_y u_d|^2 dx.$$

But from (5), u_d satisfies almost everywhere in x

$$\int_{\mathcal{O}} \nabla_y u_d(x) \nabla_y u(x) dy = \int_{\mathcal{O}} f(x) u(x) dy,$$

Then

$$\bar{J}(u) = \int_{\Omega} |\nabla u|^2 dx - 2 \int_{\Omega} f u dx + \int_{\Omega} |\nabla_y u_d|^2 dx.$$

Now it is clear that $\bar{J}(u)$ is the energy functional associated to (\mathcal{P}_0) up to a constant term. We introduce the adjoint state p by

$$\begin{cases} \frac{\partial p}{\partial x} = -\Delta_y u - f & \text{in } \Omega, \\ p(0) = 0. \end{cases}$$

Then, since $-\Delta_y u - f \in L^2(0, 1; H^{-1}(\mathcal{O}))$, we know (see Theorem 1.2 of chapter III of [4]) that $p \in H^1(0, 1; H^{-1}(\mathcal{O}))$. Furthermore, since $u \in Y$, we also deduce that $\frac{\partial p}{\partial x} \in H^{-1}(0, 1; L^2(\mathcal{O}))$ and therefore, $p \in L^2(\Omega)$. Now for every $h \in X_0$, we know that

$$\begin{aligned} \int_0^1 \langle -\Delta_y u - f, h \rangle_{H^{-1}(\mathcal{O}) \times H_0^1(\mathcal{O})} dx &= \int_0^1 \langle \frac{\partial p}{\partial x}, h \rangle_{H^{-1}(\mathcal{O}) \times H_0^1(\mathcal{O})} dx \\ &= - \int_0^1 \int_{\mathcal{O}} p \frac{\partial h}{\partial x} dx dy. \end{aligned}$$

Therefore, from optimality condition (8) we deduce that

$$\begin{aligned} \int_0^1 \langle -\Delta_y u - f, h \rangle_{H^{-1}(\mathcal{O}) \times H_0^1(\mathcal{O})} dx + \int_0^1 \int_{\mathcal{O}} \frac{\partial u}{\partial x} \frac{\partial h}{\partial x} dx dy &= \\ \int_0^1 \int_{\mathcal{O}} (-p + \frac{\partial u}{\partial x}) u dx dy &= 0, \quad \forall u \in \mathcal{U}_{ad}. \end{aligned} \quad (9)$$

Then we have obtained the optimality system

$$\begin{cases} -\frac{\partial u}{\partial x} = -p, & u(1) = u_1, \\ \frac{\partial p}{\partial x} = -\Delta_y u - f, & p(0) = 0, \end{cases}$$

which has the same associated Riccati equation (see Section 4 of chapter III of [4]) that the system of equations for Q and w of Section 2.

5. Representation formula for the solution of the Riccati equation

Let $X(x) \in \mathcal{L}(H_{00}^{1/2}(\mathcal{O}), H_{00}^{1/2}(\mathcal{O})')$, $Y(x) \in \mathcal{L}(H_{00}^{1/2}(\mathcal{O}), H_{00}^{1/2}(\mathcal{O}))$ denote

$$\begin{aligned} X(x) : \quad u(1) &\longrightarrow \frac{du}{dx}(x), \\ Y(x) : \quad u(1) &\longrightarrow u(x), \end{aligned}$$

where u is solution of (P_0) assuming $f = 0$ and $u_0 = 0$. They satisfy ($' = \frac{d}{dx}$)

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{pmatrix} 0 & -\Delta_y \\ I & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}$$

with $Y(1) = I$ and $X(0) = 0$. Furthermore Q such that $Q(x) = X(x)Y(x)^{-1}$ satisfies the differential Riccati equation $Q' + Q^2 = -\Delta_y$.

Let P_0 be the positive solution of the algebraic Riccati equation

$$-P_0\Delta_y P_0 = I, \quad P_0 = (-\Delta_y)^{-1/2}$$

By the change of variable

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} I & -P_0^{-1} \\ P_0 & I \end{pmatrix} \begin{pmatrix} \Phi \\ \Psi \end{pmatrix}$$

equations for X and Y are diagonalized in

$$\begin{pmatrix} \Phi' \\ \Psi' \end{pmatrix} = \begin{pmatrix} P_0^{-1} & 0 \\ 0 & -P_0^{-1} \end{pmatrix} \begin{pmatrix} \Phi \\ \Psi \end{pmatrix}$$

$$W = \Phi\Psi^{-1} = (X + P_0^{-1}Y)(Y + P_0X)^{-1} = (Q + P_0^{-1})(I - P_0Q)^{-1} = \mathcal{H}(Q)$$

satisfies the linear equation

$$W' = P_0^{-1}W + WP_0^{-1}$$

with $W(0) = \mathcal{H}(Q(0)) = \mathcal{H}(0) = P_0^{-1}$

$$\begin{array}{ccc} Q(0) & \longrightarrow & Q(x) \\ \downarrow \mathcal{H} & & \uparrow \mathcal{H}^{-1} \\ W(0) & \longrightarrow & W(x) \end{array}$$

$$0 \leq W \leq P_0^{-1} \Rightarrow Q = (W - P_0^{-1})(I + P_0 W)^{-1}$$

is well defined.

5.1. Dirichlet boundary condition at $x = 0$

It corresponds to a singularity of

$$Q(0) = \mathcal{H}^{-1}(W(0)) = (W(0) - P_0^{-1})(I + P_0 W(0))^{-1}$$

i.e. $W(0) = -P_0^{-1}$.

$$0 > W > -P_0^{-1} \text{ for } 0 < x \leq 1 \quad \Rightarrow \quad Q = (W - P_0^{-1})(I + P_0 W)^{-1}$$

is well defined for $0 < x \leq 1$.

6. The QR factorization

As the factorization (3) is viewed as an infinite dimensional generalization of the *LU* block triangular factorization, we now turn to the *QR* factorization, i.e. as the product of an orthogonal operator Q and an upper triangular part R . We begin by writing the normal equation for (\mathcal{P}_0) . We assume that f is regular enough and is null in a neighbourhood of Σ in order to avoid non-homogeneous boundary conditions.

$$\begin{cases} \Delta^2 u = -\Delta f & \text{in } \Omega, \\ u|_{\Sigma} = \Delta u|_{\Sigma} = 0, & -\frac{\partial u}{\partial x}|_{\Gamma_0} = u_0, \quad \frac{\partial \Delta u}{\partial x}|_{\Gamma_0} = -\frac{\partial f}{\partial x}|_{\Gamma_0} \\ u|_{\Gamma_1} = u_1, & \Delta u|_{\Gamma_1} = -f|_{\Gamma_1}. \end{cases} \quad (10)$$

Of course (\mathcal{P}_0) and (10) have the same solution. Similarly to section 2 we embed (10) in a family defined on Ω_s with additional boundary condition on Γ_s :

$$u|_{\Gamma_s} = h, \quad \Delta u|_{\Gamma_s} = k,$$

and we set

$$\frac{\partial u}{\partial x}|_{\Gamma_s} = Q(s)h + P(s)k + r(s). \quad (11)$$

If we choose $k = -f$, u is the solution of $(\mathcal{P}_{s,h})$, then Q (which should not be confused with the orthogonal part of the factorization) is the same operator as the one defined in section 2 and it satisfies (2). We also deduce that $r - Pf = w$ satisfies (2). Furthermore, from (10), Δu can be viewed as satisfying $(\mathcal{P}_{s,h})$ with right hand side Δf . Hence it

admits the following factorization

$$\begin{cases} \frac{dt}{dx} + Qt = -\Delta f, & t(0) = -\frac{\partial f}{\partial x}|_{\Gamma_0}, \\ -\frac{d\Delta u}{dx} + Q\Delta u = -t, & u(1) = -f(1). \end{cases} \quad (12)$$

Deriving (11) along a trajectory $u(x)$ and by identification we get for Q and P

$$\frac{dQ}{dx} + Q^2 + \Delta_y = 0, \quad Q(0) = 0, \quad (13)$$

$$\frac{dP}{dx} + PQ + QP = I, \quad P(0) = 0. \quad (14)$$

Both operators P and Q are self-adjoint and positive for $x > 0$. For that particular problem they commute with Δ_y . The term independent of u and Δu yields

$$Pt + \frac{dr}{dx} + Qr = 0.$$

Then, using (12), we get the decoupled form of (10) as

$$\begin{aligned} P^{-1} \frac{d^2 r}{dx^2} - (P^{-2} - 2P^{-1}Q - 2QP^{-1}) \frac{dr}{dx} \\ - (P^{-2} - 2QP^{-1}Q + P^{-1}\Delta_y) r = \Delta f, \end{aligned} \quad (15)$$

$$r(0) = -u_0, \quad \frac{\partial r}{\partial x}(0) = 0, \quad (16)$$

where the initial conditions are derived from (11) and its derivative written on Γ_0 and

$$\frac{d^2 u}{dx^2} - P^{-1} \frac{du}{dx} + (P^{-1}Q + \Delta_y) u = -P^{-1}r, \quad (17)$$

$$u(1) = u_1, \quad \frac{du}{dx}(1) = Q(1)u_1 - P(1)f(1) + P(1)r(1). \quad (18)$$

Let us denote \mathcal{Q} the mapping $f \rightarrow P^{-1}r$ defined by (15) from $Y = \{u \in L^2(\Omega) | \Delta u \in L^2(\Omega)\}$ into itself, and $\tilde{r} = P^{-1}r$. Then \tilde{r} satisfies

$$\frac{d^2 \tilde{r}}{dx^2} + P^{-1} \frac{d\tilde{r}}{dx} + (2QP^{-1} + P^{-1}Q - P^{-2} + \Delta_y) \tilde{r} = \Delta f. \quad (19)$$

Now it can be checked using (13) that the differential operators applied to \tilde{r} and u in (19), (17) respectively are adjoint of each other. Then

$$\mathcal{Q}^* \mathcal{Q} : f \rightarrow -\Delta u,$$

and \mathcal{Q} is orthogonal. Denoting \mathcal{R} the operator defined by the Cauchy problem (17) ($\mathcal{R}u = \tilde{r}$), we have obtained the factorization of (\mathcal{P}_0) as $\mathcal{Q}^*\mathcal{R}$. It can be checked that $P^{-1}\mathcal{Q} + \Delta_y$ is positive and (17) is an hyperbolic problem with damping.

7. Joint factorization for a control problem

7.1. Statement of the control problem and optimality system

Now we consider a control problem for an elliptic state equation. The state equation is

$$(\mathcal{P}_c) \begin{cases} -\Delta u = f + Bv & \text{in } \Omega, \\ u|_{\Sigma} = 0, \quad \frac{\partial u}{\partial x}|_{\Gamma_0} = -u_0, \quad \frac{\partial u}{\partial x}|_{\Gamma_1} = 0, \end{cases}$$

where the control v lies in a Hilbert space V identified to its dual and $B \in \mathcal{L}(V; L^2(\Omega))$. Define the cost function

$$J(v) = \int_{\Gamma_1} |u|_{\Gamma_1} - u_d|^2 dy + \nu \|v\|_V^2.$$

Then defining the adjoint state p by

$$\begin{cases} -\Delta p = 0 & \text{in } \Omega, \\ p|_{\Sigma} = 0, \quad \frac{\partial p}{\partial x}|_{\Gamma_0} = 0, \quad \frac{\partial p}{\partial x}|_{\Gamma_1} = u|_{\Gamma_1} - u_d, \end{cases}$$

the minimum of J is characterized by $\bar{v} = -\frac{1}{\nu} B^* p$.

7.2. Factorization of the optimality system

The factorization via dynamic programming applied to the state equation in the previous section can be applied to the coupled system of state and adjoint state equations. Setting $u|_{\Gamma_s} = \varphi$ and $p|_{\Gamma_s} = \psi$ we define a family of problems depending on s , φ , ψ by

$$\begin{cases} -\Delta u = f - \frac{1}{\nu} BB^* p & \text{in } \Omega_s, \\ u|_{\Sigma} = 0, \quad \frac{\partial u}{\partial x}|_{\Gamma_0} = 0, \quad u|_{\Gamma_s} = \varphi, \\ -\Delta p = 0 & \text{in } \Omega_s, \\ p|_{\Sigma} = 0, \quad \frac{\partial p}{\partial x}|_{\Gamma_0} = 0, \quad p|_{\Gamma_s} = \psi. \end{cases} \quad (20)$$

Then the mapping $(\varphi, \psi) \rightarrow (u, p)$ is affine but p depends linearly on ψ and not on φ . Furthermore the mapping $\psi \rightarrow \frac{\partial p}{\partial x}|_{\Gamma_s}$ as well as the mapping $\varphi \rightarrow \frac{\partial u}{\partial x}|_{\Gamma_s}$ if $\psi = 0$ are exactly the DtN mapping Q satisfying (2). So there exists a linear mapping $\overline{P}(s)$ such that

$$\begin{aligned}\frac{\partial u}{\partial x}|_{\Gamma_s} &= \bar{P}(s)\psi + Q(s)\varphi + w(s), \\ \frac{\partial p}{\partial x}|_{\Gamma_s} &= Q(s)\psi.\end{aligned}\tag{21}$$

The equation satisfied by \bar{P} is obtained in a similar fashion. Let us derive the first equation (21) along a solution of (20) and substituting the derivatives of u and p from (21)

$$\frac{\partial^2 u}{\partial x^2} = \frac{d\bar{P}}{dx}p + \bar{P}Qp + \frac{dQ}{dx}u + Q(\bar{P}p + Qu + w) + \frac{dw}{dx} = -\Delta_y u - f + \frac{1}{\nu}BB^*p.$$

Using the equation for Q from (2) and the fact that p is arbitrary we obtain the equation for \bar{P}

$$\frac{d\bar{P}}{dx} + \bar{P}Q + Q\bar{P} = \frac{1}{\nu}BB^* \quad \bar{P}(0) = 0.\tag{22}$$

$$\frac{dw}{dx} + Qw = -f, \quad w(0) = -u_0\tag{23}$$

Knowing Q from (2), equation (22) is linear. One can show that it is well posed and its solution is self-adjoint.

Now, equations for Q , \bar{P} and w being integrated once for all, for any new measurement u_d the optimal control \bar{v} is obtained in the following way

- find an initial condition for p at $x = 1$ from the system

$$\begin{aligned}\bar{P}(1)p(1) + Q(1)u(1) + w(1) &= 0, \\ Q(1)p(1) &= u(1) - u_d,\end{aligned}$$

which gives $p(1) = -(\bar{P}(1) + Q(1)^2)^{-1}(Q(1)u_d + w(1))$.

- integrate the equation for p backwards from Γ_1 to Γ_0 $\frac{dp}{dx} - Qp = 0$
- the optimal control \bar{v} is then given by $\bar{v} = -\frac{1}{\nu}B^*p$.

References

- [1] Angel E., Bellman R. *Dynamic Programming and Partial Differential Equations* Academic Press 1972.
- [2] Henry J., Ramos A. *Factorization of Second Order Elliptic Boundary Value Problems by Dynamic Programming*, submitted to Journal Math. Ana. Appl.
- [3] Henry, J., Ramos, A. *A Direct Study in a Hilbert-Schmidt Framework of the Riccati Equation Appearing in a Factorization Method of Second Order Elliptic Boundary Value Problems* Rapport de Recherche INRIA 4451.
- [4] Lions, J.L. (1968). *Contrôle Optimal de Systèmes Gouvernés par des Équations aux Dérivées Partielles*. Dunod.
- [5] Lions, J.L., and Magenes E. (1968). *Problèmes aux Limites Non Homogènes et Applications*, vol 1. Dunod.

SMOOTH MAPPINGS AND NON \mathcal{F}_T -ADAPTED SOLUTIONS ASSOCIATED WITH HAMILTON-IACOBI STOCHASTIC EQUATIONS

Daniela Ijacu

Academy of Economic Studies, Bucharest, Romania

Constantin Varsan

Institute of Mathematics, Romanian Academy

P.O. Box 1-764, RO-70700 Bucharest, Romania

cvarsan@imar.ro

Abstract Stochastic partial differential equations of Hamilton-Iacobi type including non \mathcal{F}_t -adapted solutions are studied. Using a Stratonovich type stochastic integral and an orbit solutions in a finite dimensional Lie algebra, we are dealing with non \mathcal{F}_t -adapted solutions associated with an extended characteristic system of stochastic differential equations.

Keywords: Stochastic partial differential equations, Stratonovich type stochastic integral and non \mathcal{F}_t -adapted solutions.

1. Introduction

The analysis is concentrated on stochastic partial differential equations (SPDEs) driven by the following dynamics:

$$(1) \quad \begin{cases} d_t u = g_0^\omega(x, u, \partial_x u) dt + \sum_{j=1}^m \chi_\tau(t) g_j(x, u, \partial_x u) \otimes dw_j(t) \\ u(0) = u_0^\omega(x), \quad (t, x, u) \in [0, T] \times \mathbb{R}^n \times \mathbb{R}, \\ (u, \partial_x u) \in B(0, \rho) \subseteq \mathbb{R}^{n+1} \end{cases}$$

where $u_0^\omega(\cdot) \in C_b^2(\mathbb{R}^n)$ and the continuous scalar function $g_0^\omega(\cdot) \in C_b^2(\mathbb{R}^n \times B(0, \rho))$ are only \mathcal{F} -measurable on the parameter $\omega \in \Omega$ in a complete probability space $\{\Omega, \mathcal{F}, P\}$.

Here $w(t) = (w_1, \dots, w_m(t)) \in \mathbb{R}^m$, $t \in [0, T]$, is a standard m -dimensional Wiener process and $\tau(\omega) : \Omega \rightarrow [0, T]$ is a stopping time allowing

one to define non \mathcal{F}_t -adapted and bounded solutions $(u, \partial_x u) \in B(0, \rho) \subseteq \mathbb{R}^{n+1}$ on a complete filtered probability space $\{\Omega, \mathcal{F}, P; \{\mathcal{F}_t\} \nearrow \mathcal{F}\}$. The stochastic integral “ \otimes ” appearing in the equation (1) coincides with the usual Fisk-Stratonovich integral “ \circlearrowright ” provided the dependence on $\omega \in \Omega$ of $u_0(\cdot)$ and $g_0(\cdot)$ is omitted.

As far as the given $u_0^\omega(\cdot)$ and $g_0^\omega(\cdot)$ are not \mathcal{F}_t -adapted we need to define a special type of stochastic integral “ \otimes ” (Stratonovich type) encompassing non \mathcal{F}_t -adapted solutions $(u(t, x, \omega), \partial_x u(t, x, \omega))$, $t \in [0, T]$ and fulfilling (1) along to the corresponding trajectories $x = \tilde{x}(t, \lambda, \omega)$, $t \in [0, T]$ contained in the characteristic system. It can be accomplished using the Langevin’s approximation $w^\varepsilon(t)$, $t \in [0, T]$, $\varepsilon \in (0, 1]$ of the original Wiener process and an adequate rule of stochastic differentiation. From now on we shall not mention the explicit dependence on $\omega \in \Omega$ and write $u_0(\cdot)$, $g_0(\cdot)$.

On the other hand, the dependence on the gradient $\partial_x u \in \mathbb{R}^n$ of the given $g_i(\cdot)$, $i \in \{0, 1, \dots, m\}$, is obstructing the martingale approach and to achieve the goal we use some smooth mappings generated by the finite dimensional Lie algebra $L(Z_1, \dots, Z_m)$ associated with the given smooth diffusion coefficients $\{g_1, \dots, g_m\}$.

A motivation for considering such problems may appear from describing a non \mathcal{F}_t -adapted solutions fulfilling the corresponding Hamilton-Iacobi stochastic differential equations associated with an optimal solution $(\tilde{x}(t, \omega), \tilde{u}(t, \omega)$, $t \in [0, T]$; they are based on an augmented Lagrangean $H(t, x, u, \psi; dt, dw(t))$ defined as a stochastic differential form

$$H(t, x, u, \psi; dt, dw(t)) = [\psi f(t, x, u) + f_0(t, x, u)]dt + \sum_{j=1}^m \psi g_j(t, x) \otimes dw_j(t)$$

Here the adjoint row vector function $\psi = \tilde{\psi}(t, \omega)$ has to be determined as a solution of a stochastic differential equation

$$(2) \quad \begin{cases} dt\tilde{\psi} = -\frac{\partial H}{\partial x}(t, \tilde{x}(t, \omega), \tilde{u}(t, \omega), \tilde{\psi}; dt, dw(t)), & t \in [0, T] \\ \tilde{\psi}(t_f, \omega) = \partial_x F(\tilde{x}(T, \omega)) \end{cases}$$

and the optimal pair $(\tilde{x}(t, \omega), \tilde{u}(t, \omega))$ obeys to

$$(3) \quad \begin{cases} dt\tilde{x} = \frac{\partial H}{\partial \psi}(t, \tilde{x}, \tilde{u}(t, \omega), \tilde{\psi}(t, \omega); dt, dw(t)), & t \in [0, T] \\ \tilde{x}(0) = x_0 \in X \subseteq \mathbb{R}^n \end{cases}$$

$$(4) \quad \begin{cases} \min_{u \in U \subseteq R^k} \tilde{\psi}(t, \omega) f(t, \tilde{x}(t, \omega), u) + f_0(t, \tilde{x}(t, \omega), u) = \\ \tilde{\psi}(t, \omega) f(t, \tilde{x}(t, \omega), \tilde{u}(t, \omega)) + f_0(t, \tilde{x}(t, \omega), \tilde{u}(t, \omega)) \\ \text{a.e. } (t, \omega) \in [0, T] \times \Omega \end{cases}$$

Using a smooth mapping $x=G(p, \hat{x})$, $\hat{x} \in B(x_0, \rho_0) \subseteq \mathbb{R}^n$, $p \in B(0, \rho) \subseteq \mathbb{R}^M$, we decompose the non \mathcal{F}_t -adapted solutions $(\tilde{x}(t, \omega), \psi(t, \omega))$, $t \in [0, T]$ into a continuous and \mathcal{F}_t -adapted process valued in the place of smooth diffeomorphisms and a corresponding continuously differentiable process $(\tilde{x}(t, \omega), \tilde{\psi}(t, \omega))$, $t \in [0, T]$ such that (2) and (3) are satisfied provided an adequate stochastic integral “ \otimes ” is defined.

2. Some auxiliary lemmas and main results

Everywhere in this paper we assume that smooth deterministic functions $g_j(x, u, p)$ are given such that $g_j \in C_b^\infty(\mathbb{R}^n \times B(0, \rho))$, $j \in \{1, \dots, m\}$ where the ball $B(0, \rho) \subseteq \mathbb{R}^{n+1}$ is fixed. Denote $z = (u, p, x) \in \mathbb{R}^{2n+1}$, $D \stackrel{\text{def}}{=} B(0, \rho) \times \mathbb{R}^n$ and define smooth vector fields $Z_j(z) = \begin{pmatrix} Y_j(z) \\ X_j(z) \end{pmatrix}$, with $X_j(z) = -\partial_p g_j(x, u, p)$ in \mathbb{R}^n , $Y_j(z) = \begin{pmatrix} g_j(x, u, p) - \langle p, \partial_p g_j(x, u, p) \rangle \\ \partial_x g_j(x, u, p) + p \partial_u g_j(x, u, p) \end{pmatrix}$ in \mathbb{R}^{n+1} . A solution for SPDEs (1) is derived using the corresponding stochastic system of characteristics.

In addition we have to start with a local solution associated with the reduced stochastic differential system:

$$(5) \quad d_t z = \sum_{j=1}^m \chi_\tau(t) Z_j(z) \odot dw_j(t), \quad t \in [0, T], \quad z \in D = B(0, \rho) \times \mathbb{R}^n$$

$$z(0) = z_0 \in D_0 = B(0, \rho_0) \times \mathbb{R}^n, \quad 0 < \rho_0 < \rho$$

where the Fisk-Stratonovich integral “ \odot ” is used and $w(t) = (w_1(t), w_2(t), \dots, (w_m(t)) \in \mathbb{R}^m$ is a standard m -dimensional Wiener process on a given filtered probability space $(\Omega, \mathcal{F}, P; \{\mathcal{F}_t\} \nearrow \mathcal{F})$.

A local solution fulfilling (5) is found as a continuous and \mathcal{F}_t -adapted process valued in the space of smooth mappings $z \in C_b^\infty(D_0; \mathbb{R}^{2n+1})$ and it is done assuming

(\mathcal{H}) The Lie algebra $L(Z_1, \dots, Z_m) \subseteq C_b^\infty(D_0; \mathbb{R}^{2n+1})$ determined by the vector fields $\{Z_1, \dots, Z_m\}$ is finite dimensional.

The assumption (\mathcal{H}) allow us to fixe a system of generators $\{Z_1, \dots, Z_m, Z_{m+1}, \dots, Z_M\} \subseteq L(Z_1, \dots, Z_m)$ and to define the corresponding orbit of smooth mappings.

$$(6) \quad \begin{cases} S(p, z_0) = S_1(t_1) \odot \dots \odot S_M(t_M)(z_0) \\ p = (t_1, \dots, t_M) \in D_M = \prod_{j=1}^M [-a_j, a_j], \text{ for } z_0 \in D_0 = B(0, \rho_0) \times \mathbb{R}^n \end{cases}$$

where $S_j(t, z_0)$, $t \in [-a_j, a_j]$, $z_0 \in D_0$, is a local flow generated by the vector field Z_j , $j \in \{1, \dots, M\}$.

Using the nonsingular algebraic representation of the associated gradient system given in (7) we are able to recover the original vector fields $\{Z_1, \dots, Z_M\}$ along to the orbit solution (6) and some analitic vector fields $q_j \in A(D_M; \mathbb{R}^M)$, $j \in \{1, \dots, M\}$, are defined such that

$$(7) \quad \begin{cases} \frac{\partial S}{\partial p}(p, z_0) q_j(p) = Z_j(S(p, z_0)) & j \in \{1, \dots, M\} \text{ } p \in D_M, \text{ } z_0 \in D_0 \\ \text{the } (M \times M) \text{ matrix } Q(p) = (q_1(p), \dots, q_M(p)), \text{ } p \in D_M \\ \text{is a non singular one} \end{cases}$$

A local solution for the stochastic differential system (5) is constructed using the mapping $S(p, z_0)$ in (6) provided an \mathcal{F}_t -adapted continuous process $p = p(t) \in D_M$, $t \in [0, T]$ is defined as a solution of the following system:

$$(8) \quad d_t p = \sum_{j=1}^m \alpha(p) q_j(p) \odot dw_j(t), \text{ } p(0) = 0, \text{ } p \in \mathbb{R}^M$$

where the smooth scalar function $\alpha \in C^\infty(\mathbb{R}^M; [0, 1])$ is taken adequately and fulfilling $\alpha(p) = 0$ for $p \in \mathbb{R}^M \setminus B(0, 2\hat{\rho})$, $\alpha(p) = 1$ for $p \in B(0, \hat{\rho})$, where $\hat{\rho} > 0$ is fixed such that $B(0, 2\hat{\rho}) \subseteq D_M$. Let $\tau(\omega) : \Omega \longrightarrow [0, T]$ be a stopping time defined by $\tau(\omega) = \inf\{t \in [0, T]; |p(t)| > \hat{\rho}\}$ and associated with the solution $p = p(t)$, $t \in [0, T]$, globally defined in (8). It is easily seen that $\hat{p}(t) = p(t \wedge \tau) \in B(0, \hat{\rho})$, $t \in [0, T]$, is obeying to the following stochastic system:

$$\hat{p}(t) = \sum_{j=1}^m \int_0^t \chi_\tau(s) q_j(\hat{p}(s)) \odot dw_j(s) \text{ } \hat{p}(0) = 0, \text{ } t \in [0, T]$$

where $\chi_\tau(t) = 1$ for $\tau > t$ and $\chi_\tau(t) = 0$ for $\tau \leq t$, $t \in [0, T]$.

The \mathcal{F}_t -adapted and continuous process

$$z(t, z_0) = S(p(t), z_0), \text{ } t \in [0, T], \text{ } z_0 \in D_0 = B(0, \rho_0) \times \mathbb{R}^n$$

will be a local solution for the stochastic system in (5) fulfilling the following integral equations:

$$Z(t, z_0) = z_0 + \sum_{j=1}^m \int_0^t \chi_\tau(s) Z_j(z(s, z_0)) \odot dw_j(s) \text{ } t \in [0, T], \text{ } z_0 \in D_0$$

A local solution for SPDEs (1) can be constructed provided a continuously differentiable process $z_0 = z_0(t, \lambda) \in D_0$, $t \in (0, a]$, $0 < a \leq T$, is defined such that

$$(9) \quad z(t, \lambda) = S(p(t); z_0(t, \lambda)) \in D, \text{ } t \in [0, a], \text{ } \lambda \in \mathbb{R}^n$$

is a local solution of an extended system of characteristics

$$(10) \quad \begin{cases} d_t z = Z_0(z) dt + \sum_{j=1}^m \chi_\tau(t) Z_j(z) \otimes dw_j, & t \in [0, T], \quad z \in B(0, \rho) \times \mathbb{R}^n \\ z(0) = z_0(\lambda) = (u_0(\lambda), \partial_\lambda u_0(\lambda)) \in B(0, \rho_1) \times \mathbb{R}^n = D_1, & 0 < \rho_1 < \rho \end{cases}$$

where $z = (u, \partial_x u, x) = (u, p, x)$ and the smooth vector field $Z_0(z) = \begin{pmatrix} Y_0(z) \\ X_0(z) \end{pmatrix}$ is associated with the drift part g_0 in (1) as follows:

$$(11) \quad \begin{cases} X_0(z) = -\partial_p g_0(x, u, p) \in \mathbb{R}^n, \\ Y_0(z) = \left(g_0(x, u, p) + \langle p, X_0(z) \rangle \right. \\ \left. \partial_x g_0(x, u, p) + p \partial_u g_0(x, u, p) \right) \end{cases}$$

In addition the Stratonovich type integral “ \otimes ” is computed passing to the limit $\varepsilon \searrow 0$ in an ordinary rule of derivation applied to the smooth mapping

$$(12) \quad z^\varepsilon(t, \lambda) = S(p^\varepsilon(t); z_0(t, \lambda)), \quad t \in [t', t''] \subset (0, a], \quad 0 < a \leq T$$

where $p = p^\varepsilon(t)$, $t \in [0, T]$, is fulfilling the following system of ordinary differential equations:

$$\frac{dp}{dt} = \sum_{j=1}^m \chi_\tau(t) \alpha(p) q_j(p) \frac{dw_j^\varepsilon(t)}{dt}, \quad t \in [0, T], \quad p(0) = 0$$

provided the Langevin's smooth approximation

$$w^\varepsilon(t) = \int_0^t y^\varepsilon(s) ds = w(t) - \eta(t, \varepsilon), \quad d_t y^\varepsilon(t) = -\frac{1}{\varepsilon} y^\varepsilon(t) dt + \frac{1}{\varepsilon} dw(t)$$

for $t \in [0, T]$, $\varepsilon \in (0, 1]$, is used.

As a consequence we may and do write the following

Definition 1 A stochastic integral “ \otimes ” appearing in (10) is computed as follows:

$$\begin{aligned} \int_{t'}^{t''} \chi_\tau(t) Z_j(z(t, \lambda)) \otimes dw_j(t) &= \left[\int_{t'}^{t''} \chi_\tau(t) Z_j(S(\hat{p}(t); z_0)) \circ dw_j(t) \right]_{z_0=z_0(t', \lambda)} + \\ &\quad \int_{t'}^{t''} \left[\int_t^{t''} \frac{\partial}{\partial z_0} (\chi_\tau(\sigma) Z_j(S(\hat{p}(\sigma); z_0)) \circ dw_j(\sigma) \right]_{z_0=z_0(t, \lambda)} \frac{dz_0}{dt}(t, \lambda) dt \end{aligned}$$

$j \in \{1, \dots, m\}$, where the Fisk-Stratonovich stochastic integral “ \circ ” associated with continuous \mathcal{F}_t -adapted process valued in the space of smooth mappings $C^\infty(D_0; \mathbb{R}^{2n+1})$ is used.

Definition 2 Let $\varphi \in C_b^{0,3}([0, a] \times D; \mathbb{R})$ and $\hat{z}(t, \lambda) = S(\hat{p}(t))$, $z_0(t, \lambda) \in D$, $t \in [0, a]$, $\lambda \in \mathbb{R}^n$ be defined as in (9). Then

$$\begin{aligned} & \int_{t'}^{t''} \chi_\tau(t) \varphi(t, \hat{z}(t, \lambda)) \otimes dw_j(t) = \\ & \quad \left[\int_{t'}^{t''} \chi_\tau(t) \varphi(t, S(\hat{p}(t); z_0)) \odot dw_j(t) \right]_{z_0=z_0(t', \lambda)} + \\ & \quad \int_{t'}^{t''} \left[\int_t^{t''} \frac{\partial}{\partial z_0} (\chi_\tau(\sigma) \varphi(\sigma, S(\hat{p}(\sigma); z_0))) \odot dw_j(\sigma) \right]_{z_0=z_0(t, \lambda)} \frac{dz_0}{dt}(t, \lambda) dt \end{aligned}$$

The following stochastic rule of derivation holds true.

Lemma 1 Let $f \in C_b^{1,3}([0, T] \times D; \mathbb{R})$ be given and consider the solution $\hat{z}(t, \lambda) = S(\hat{p}(t); z_0(t, \lambda))$, $t \in [0, a]$, fulfilling the following integral equations

$$\begin{aligned} \hat{z}(t'', \lambda) - \hat{z}(t', \lambda) &= \int_{t'}^{t''} \frac{\partial S}{\partial z_0}(\hat{p}(t); z_0(t, \lambda)) \frac{dz_0}{dt}(t, \lambda) dt + \\ & \quad \sum_{j=1}^m \int_{t'}^{t''} \chi_\tau(t) Z_j(\hat{z}(t, \lambda)) \otimes dw_j(t), \quad [t', t''] \subseteq [0, a] \end{aligned}$$

Then $E \stackrel{\text{def}}{=} f(t'', \hat{z}(t'', \lambda)) - f(t', \hat{z}(t', \lambda)) \stackrel{\text{def}}{=} \lim_{\varepsilon \rightarrow 0} \int_{t'}^{t''} \frac{d}{dt} [f(t, z^\varepsilon(t, \lambda))] dt$ can be expressed as follows

$$\begin{aligned} E &= \int_{t'}^{t''} [\partial_t f(t, \hat{z}(t, \lambda)) + \langle \partial_z f(t, \hat{z}(t, \lambda)), \partial_{z_0} S(\hat{p}(t); z_0(t, \lambda)) \frac{dz_0}{dt}(t, \lambda) \rangle] dt + \\ & \quad \sum_{j=1}^m \int_{t'}^{t''} \chi_\tau(t) \langle \partial_z f(t, \hat{z}(t, \lambda)), Z_j(\hat{z}(t, \lambda)) \rangle \otimes dw_j(t) \end{aligned}$$

where the stochastic integral “ \otimes ” is acting as in the Definition 2, and $z^\varepsilon(t, \lambda) = S(p^\varepsilon(t); z_0(t, \lambda))$ is defined in (12).

Relying on the integral equations given in Lemma 1 we may and do choose $z_0 = z_0(t, \lambda)$ such that the characteristic system in (10) is fulfilled. In this respect, let $z_0 = z_0(t, \lambda)$, $t \in [0, a]$, $0 < a \leq T$, be the unique solution of the following system of ordinary differential equations

$$(13) \quad \begin{cases} \frac{dz_0}{dt}(t, \lambda) = \left[\frac{\partial S}{\partial z_0}(\hat{p}(t); z_0(t, \lambda)) \right]^{-1} Z_0(S(\hat{p}(t); z_0(t, \lambda))) \\ z_0(0, \lambda) = z_0(\lambda) = (y_0(\lambda); \lambda) \in B(0, \rho_1) \times \mathbb{R}^n, \quad 0 < \rho_1 < \rho \end{cases}$$

where the vector field $Z_0 \in C_b^1(D, \mathbb{R}^{2n+1})$ is defined in (11). By a direct computation we get the following

Lemma 2 Assume the hypothesis (\mathcal{H}) is fulfilled and define $\hat{z}(t, \lambda) = S(\hat{p}(t); z_0(t, \lambda))$, where $z_0(t, \lambda), t \in [0, a], \lambda \in \mathbb{R}^n$ is the unique solution associated with (13). Then $z = \hat{z}(t, \lambda), t \in [0, a], \lambda \in \mathbb{R}^n$, is a local solution of (10) obeying to

$$\hat{z}(t, \lambda) = z_0(\lambda) + \int_0^t Z_0(\hat{z}(s, \lambda)) ds + \sum_{j=1} \int_0^t \chi_\tau(s) Z_j(\hat{z}(s, \lambda)) \otimes dw_j(s), \quad \forall t \in [0, a], \lambda \in \mathbb{R}^n$$

By definition $\hat{z}(t, \lambda) = (\hat{y}(t, \lambda); \hat{x}(t, \lambda))$ obey to the integral equation in Lemma 2. We may and do solve the following algebraic equations:
 $\hat{x}(t, \lambda) = x$ and find $\lambda = \psi(t, x)$, $t \in [0, a]$, $x \in \mathbb{R}^n$ such that

$$(14) \quad \hat{x}(t, \psi(t, x)) = x, \quad \psi(t, \hat{x}(t, \lambda)) = \lambda, \quad t \in [0, a], \quad x \in \mathbb{R}^n$$

Denote $y(t, x) = \hat{y}(t, \psi(t, x)) = (u(t, x), p(t, x))$, $t \in [0, a]$, $x \in \mathbb{R}^n$ and one sees easily that

$$(15) \quad u(t, \hat{x}(t, \lambda)) = \hat{u}(t, \lambda), \quad p(t, \hat{x}(t, \lambda)) = \hat{p}(t, \lambda)$$

A local solution for SPDEs (1) is assimilated with the above given continuous process $u=u(t, x)$, $t \in [0, a]$, $\lambda, x \in \mathbb{R}^n$ provided we are able to show

$$(16) \quad \partial_x u(t, x) = p(t, x), \quad t \in [0, a], \quad x \in \mathbb{R}^n$$

and the stochastic differential $[d_t u(t, x)]_{x=\hat{x}(t, \lambda)}$ along to $x = \hat{x}(t, \lambda)$ is acting as the following integral shows:

$$(17) \quad \int_{t'}^{t''} [d_t u(t, x)]_{x=\hat{x}}(t, \lambda) = \int_{t'}^{t''} [d_t \hat{u}(t, \lambda) - \langle \hat{p}(t, \lambda), d_t \hat{x}(t, \lambda) \rangle],$$

for any $[t', t''] \subseteq [0, a]$, where the left hand side in (17) is defined as

$$(18) \quad \int_{t'}^{t''} [d_t u(t, x)]_{x=\hat{x}}(t, \lambda) = \lim_{\varepsilon \rightarrow 0} \int_{t'}^{t''} [d_t u^\varepsilon(t, x)]_{x=x^\varepsilon(t, \lambda)} dt$$

Here $u^\varepsilon(t, x) = u^\varepsilon(t, \psi^\varepsilon(t, x))$, and the smooth approximation $z^\varepsilon(t, \lambda) = (y^\varepsilon(t, \lambda), x^\varepsilon(t, \lambda)) = S(p^\varepsilon(t); z_0(t, \lambda))$ is a continuously differentiable mapping with respect to both variables $t \in [0, a]$, $\lambda \in \mathbb{R}^n$, and $\lambda = \psi^\varepsilon(t, x)$ is the unique solution of the algebraic equations

$$(19) \quad x^\varepsilon(t, \lambda) = x \in \mathbb{R}^n, \quad \psi^\varepsilon(t, x^\varepsilon(t, \lambda)) = \lambda, \quad t \in [0, a]$$

Using the continuously differentiable process $z^\varepsilon(t, \lambda)$ we get the equations (16) and (17) fulfilled and expressed as follows

Lemma 3 Under the same conditions as in Lemma 2 define $y(t, x) = y(t, \psi(t, x)) = (u(t, x), p(t, x))$, $t \in [0, a]$, $x \in \mathbb{R}^n$ as in (15). Then (16) and (17) hold true, i.e.

$$\partial_x u(t, x) = p(t, x), \quad t \in [0, a], \quad x \in \mathbb{R}^n$$

$$\int_{t'}^{t''} [d_t u(t, x)]_{x=\hat{x}(t, \lambda)} = \int_{t'}^{t''} d_t \hat{u}(t, \lambda) + \int_{t'}^{t''} \langle \hat{p}(t, \lambda), \partial_p g_0(\hat{z}(t, \lambda)) \rangle dt + \sum_{j=1}^m \int_{t'}^{t''} \chi_\tau(t) \langle \hat{p}(t, \lambda), \partial_p g_j(\hat{z}(t, \lambda)) \rangle \otimes dw_j(t)$$

for any $[t', t''] \subseteq [0, a]$, where the left hand side is given in (18).

The following theorem is a direct consequence of the results stated in Lemmas 2 and 3.

Theorem 1 Let $g_i(x, u, p)$, $i \in \{0, 1, \dots, m\}$, be given such that the hypothesis (\mathcal{H}) is fulfilled. Let $\hat{z}(t, \lambda) = (\hat{y}(t, \lambda), \hat{x}(t, \lambda))$, $(t, \lambda) \in [0, a] \times \mathbb{R}^n$, be the local solution associated with the integral equations given in Lemma 2. Let $u(t, x) = \hat{u}(t, \psi(t, x))$ and $p(t, x) = \hat{p}(t, \psi(t, x))$ where $\hat{y}(t, \lambda) = (\hat{u}(t, \lambda), \hat{p}(t, \lambda))$ and $\lambda = \psi(t, x)$ is the unique solution fulfilling (14).

Then $\partial_x u(t, x) = p(t, x)$ and $u = u(t, x)$ is a local solution of the SPDE (1) along to $x = \hat{x}(t, \lambda)$, i.e. $u(0, x) = u_0(x)$, $x \in \mathbb{R}^n$, and

$$[d_t u(t, x)]_{x=\hat{x}(t, \lambda)} = g_0(\hat{z}(t, \lambda))dt + \sum_{j=1}^m \chi_\tau(t) g_j(\hat{z}(t, \lambda)) \otimes dw_j(t)$$

for any $t \in [0, a]$, $\lambda \in \mathbb{R}^n$, where

$$\int_{t'}^{t''} [d_t u(t, x)]_{x=\hat{x}(t, \lambda)} = \hat{u}(t'', \lambda) - \hat{u}(t', \lambda) + \int_{t'}^{t''} \langle \hat{p}(t, \lambda), \partial_p g_0(\hat{z}(t, \lambda)) \rangle dt + \sum_{j=1}^m \int_{t'}^{t''} \chi_\tau(t) \langle \hat{p}(t, \lambda), \partial_p g_j(\hat{z}(t, \lambda)) \rangle \otimes dw_j(t), \quad \text{for any } [t', t''] \subseteq [0, a].$$

Comment. A SPDE of parabolic type is obtained from the equation (1) replacing the drift g_0 by $[\Delta_x u + f(x, u, \partial_x u)]$ where the Laplacian $\Delta_x u$ has to be computed along to the continuous process $x = \hat{x}(t, \lambda)$ which may involve new difficulties unless we assume, in addition, $\partial_p g_i(x, u, p) = b_i \in \mathbb{R}^n$, $i \in \{1, 2, \dots, m\}$.

References

- [1] B. Iftimie, C. Varsan, *A pathwise solution for nonlinear parabolic equations with stochastic perturbations*, (to appear).
- [2] C. Varsan, *On evolution system of differential equations with stochastic perturbations*, Preprint IMAR nr.4/2001.
- [3] C. Varsan, C. Sburlan, *Basics of mathematical physics equations and elements of differential equations* (in romanian) Ed. Ex Ponto, 2000.
- [4] C. Varsan, *Applications of Lie algebras to hyperbolic and stochastic differential equations*, Kluwer Academic Publishers, 1999.

ON THE SOBOLEV BOUNDARY VALUE PROBLEM WITH SINGULAR AND REGULARIZED BOUNDARY CONDITIONS FOR ELLIPTIC EQUATIONS

Nicolae Jitarașu

Moldova State University

60, A.Mateevici Str.

MD-2009 Chișinău, Republic of Moldova

jitarasu@usm.md

Abstract In the domain G bounded by the exterior $(n-1)$ -dimensional boundary Γ_0 and interior smooth n_k -dimensional ($0 \leq n_k \leq n-1$) manifold Γ_k ($k = 1, \dots, \chi$) without bord the boundary value problem (BVP) with boundary conditions (BC) on Γ_k is considered. For solutions $u(x)$ with power singularity on Γ_k of the Dirichlet problem for Laplace operator the asymptotical reprezentation of $u(x)$ near Γ_k is obtained. Basing on this reprezentation singular boundary condition on the Γ_k is formulated, the integral reprezentation of the solution $u(x)$ of the Sobolev BVP with singular BVC on the Γ_k $k = 1, \dots, \chi$) is obtained.

Keywords: Sobolev boundary problem, elliptic equation, singular and regularized boundary conditions.

Introduction

Let $G_0 \subset R^n$ be a bounded domain with $(n-1)$ -dimensional boundary $\Gamma_0 \in C^\infty$ and the n_k -dimensional manofolds Γ_k without boundary lying inside of Γ_0 , $0 \leq n_k \leq n-1$. Assume that $\Gamma_k \in C^\infty$, ($k = 1, \dots, \chi$) and $\Gamma_k \cap \Gamma_j = \emptyset$ for $k \neq j$. Denote by G the domain $G_0 \setminus \bigcup_{j=1}^{\chi} \Gamma_j$,

$$\Gamma = \bigcup_{j=0}^{\chi} \Gamma_j \text{ the boundary of domain } G.$$

The Sobolev boundary value problem in the classical formullation [1,2] is a boundary value problem in the domain G with differential boundary

conditions on Γ_k ($k = 0, \dots, \chi$):

$$L(x, D)u(x) = f(x) \quad (x \in G, \text{ord}L = 2m), \quad (1)$$

$$B_{j0}(x, D)u(x) |_{\Gamma_0} = \varphi_{j0}(x), \quad j = 1, \dots, m, \quad \text{ord}B_{j0} = m_j, \quad (2)$$

$$B_{jk}(x, D)u(x) |_{\Gamma_k} = \varphi_{jk}(x), \quad j = 1, \dots, q_k, \quad k = 1, \dots, \chi \quad (3)$$

The boundary conditions on the Γ_k is understood as equalities of traces $B_{jk}u |_{\Gamma_k}$ to given functions $\varphi_{jk}, D = (\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n})$.

In such formulation the number of boundary conditions (3) on each variety Γ_k depends on the degree of smoothness of the solution $u(x)$ near Γ_k . This problem in the spaces $H_2^s(G)$ of sufficiently smooth solutions has been investigated in [1,2]. The Sobolev problem in the Banach spaces $H_2^s(G)$ of distributions was studied in [3,4]. Using the Green formulas, the autors introduced the notion of generalized solution of the Sobolev problem, by defining the functional spaces $\tilde{H}_p^s(G)$ of Sobolev type for solutions and proving the completeness collection of izomorphisms generated by the operator of boundary value problem. It has been proved that for $\forall s \in R^1$ the operator of Sobolev problem realizes the izomorphism between the spases $\tilde{H}_p^s(G)$ (of solutions) and $\tilde{K}_p^s(G, \Gamma)$ (of compatible right hand sides).

In the theory mentioned above the boundary value problem (1)–(3) was studied in a closed domain \bar{G} , since the operator A of the boundary value problem in a open domain has nontrivial kernel. The dimension of the kernel $\ker A$ depends on the degrees of smoothnes of the solution $u(x)$ near and on the varieties Γ_k .

In the constructed theory the boundary conditions (3) have sense as traces on Γ_k and for smooth solutions overdetermine the boundary problem (1),(2) in \bar{G} . Those conditions (3) which lose sens as traces on Γ_k , by no means are not conected with the comportament of the solution $u(x)$ near Γ_k . However, in many problems from applied sciences the necesity for studing of solutions, having pover singularities on manifolds of various dimensions [6,7], arises.

In this case is necessary to give on Γ_k other natural conditions. These conditions are concerned either with the asymptotic behavior of the solution $u(x)$ near the manifold Γ_k or with giving on Γ_k of a regularized solution. For simplicity here we consider the Dirichlet problem in the half-space R_+^n for Laplace operator with singular boundary conditions on the variety $R^q \subset R_+^n$ and the boundary value problem with singular boundary conditions on Γ_k , $k = 1, \dots, \chi$.

The Dirichlet problem in R_+^n/R^q .

Let $x=(x_1, \dots, x_n) \in R^n$, $n \geq 3$, and let $G_0 = R_+^n = \{x \in R^n : x_n > 0\}$ be the half-space $x_n > 0$ in R^n , $x = (x', x'')$, $x' \in R^q$, $x'' \in R^{n-q}$, where

R^{n-q} is the $(n-q)$ -dimensional subspace, orthogonal to R^q in R^n . Denote by $H^S(G)$, $H^S(\Gamma)$ the Sobolev spaces in G and Γ respectively, $s \in R^1$, with norms $\|\cdot, G\|_s$, and $\|\cdot, \Gamma_1\|_s$.

Consider the following problem. Find the solution $u(x) \in H^s(G)$ of the equation

$$\Delta u(x) = f(x) \in H^{s-2}(G), \quad (4)$$

under the boundary condition on Γ_0 :

$$u(x) |_{\Gamma_0} = \varphi(x) \in H^{s-1/2}(\Gamma_0). \quad (5)$$

Evidently the solution of the problem (4),(5) in G must be defined as the restriction on G of the corresponding solution of the Dirichlet problem (4),(5) in \overline{G} for arbitrary $\overline{f}(x)$. If $\overline{f}(x)$ is an expansion of $f(x)$ on G_0 , then the solution $\overline{u}(x)$ of the Dirichlet problem (4),(5) in \overline{G} with $\overline{f}(x)$ exists and

$$\|\overline{u}(x), \overline{G}\|_s < C \left(\|\overline{f}(x), \overline{G}\|_{s-2} + \|\varphi(x), \Gamma_0\|_{s-1/2} \right)$$

with the constant C not depending of $\overline{f}(x)$. Therefore, the general solution of the problem (4),(5) in G has the forme $\overline{u}(x) = u_0(x) + u'(x)$, $u_0(x)$ and $u'(x)$ are the solutions of the Dirichlet problem (4),(5) for given $\overline{f}_0(x)$ in \overline{G} and $f'(x)$ in \overline{G} with $\text{supp } f'(x) \subset R^q$. Usually $\overline{f}_0(x)$ is a sufficiently smooth expansion of $f(x)$ from G to \overline{G} , $f'(x)$ is a singular distribution. In particular, $u'(x)$ is a solution of problem (4),(5) with $f(x) \equiv 0$ in G , $\varphi \equiv 0$ on Γ_0 . It is known [5] that the distribution $f'(x)$ with $\text{supp } f'(x) \subset R^q$ has the form

$$f'(x) = \sum_{\sigma} f_{\sigma}(x') \times D_{x''}^{\sigma} \delta(x'' - x''_0), \quad (6)$$

where $\delta(x'' - x''_0)$ is the Dirac measure concentrated on the R^q and the symbol \times means the direct product of distributions.

For sufficiently smooth $f_0(x)$ and $\varphi(x)$ the solution $u_0(x)$ of the Dirichlet problem (1),(2) in \overline{G} can be reprezented with the help of the Green function $\mathcal{G}(x, y)$ by formulae [9]

$$u(x) = \int_{R_+^n} \mathcal{G}(x, y) f_0(y) dy - \int_{\Gamma_0} \frac{\partial}{\partial \nu_y} \mathcal{G}(x, y') \varphi(y') dy', \quad y' \in R^{n-1}. \quad (7)$$

From this equality in a formal way we obtain

$$u'(x) = \sum_{\sigma} \int_{R_+^n} D_{x''}^{\sigma} \mathcal{G}(x', y', x'', x''_0) f_{\sigma}(y') dy', \quad (8)$$

where $dy' = dy_1 \cdots dy_q$, $D_{x''}^\sigma$ is the derivative with respect to the variables x'' from the orthogonal subspace R^{n-q} .

For simplicity of the analytical verification, at first we consider only term of the function (6) with $\sigma = 0$: $(f'(x) = \rho(x') \times \delta(x'' - x_0'')$. Then

$$\begin{aligned} u'(x) &= \int_{R^q} E_n(x' - y', x'' - x_0'') \rho(y') dy' + \\ &+ \int_{R^q} g_n(x, y', \bar{x}_0'') \rho(y') dy' = I_1 + I_2. \end{aligned} \quad (9)$$

Here and in what follows

$$E_n(x', x'') = (2 - n)\Omega_n^{-1}r^{2-n} \equiv a_n r^{2-n}, \quad r^2 = |x' - y'|^2 + |x'' - x_0''|^2,$$

$g_n(x, y', x_0'') = a_n \tilde{r}^{2-n}$, where $\tilde{r}^2 = |x' - y'|^2 + |x'' - \bar{x}_0''|^2$; $x_0'' = (x_{q+1}^0, \dots, x_{n-1}^0, -x_n^0)$, $\Omega_n = 2\pi^{n/2}\Gamma(\frac{n}{2})$ is the area of the surface of the sphere $|x| = 1$ in R^n , Γ is the Euler's function. Now we pass to study the behaviour of the solution $u'(x)$ near the manifold $\Gamma_1 = R^q$. For this it is necessary to introduce some notations.

Let $n \geq 3$, $n - q \leq 2$, $\rho(x') \in C_0^{n-q-1}(R^q)$. Denote by $\alpha = n - q - 2$,

$$P_\alpha(x', z')\rho = \sum_{\lambda=0}^{\alpha} \sum_{|k|=\lambda} \frac{\rho^{(k)}(x')}{k!} (-z')^k \equiv \sum_{\lambda=0}^{\alpha} \mathcal{P}_\lambda(x', z')\rho$$

is the segment of the Taylor expansion of the function $\rho(x' - z')$ at the point $z' = 0$,

$$\begin{aligned} Q_\lambda(D')\rho(x') &= \sum_{|k|=\lambda} A_k \frac{f^{(k)}(x')}{k!}, \quad A_k = \int_{|\omega'|=1} (-\omega')^k d\omega', \\ I_{11}(x) &= \int_{R^q} E_n(z', x'' - x_0'') [\rho(x' - z') - P_{\alpha-1}(x', z')\rho - \\ &- \theta(z')\mathcal{P}_\alpha(x', z')\rho] dz', \end{aligned} \quad (10)$$

where $k = (k_1, \dots, k_q)$, $D' = (\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_q})$, $\theta(z')$ is the characteristic function of the ball $B_1 = \{z' : |z'| < 1\}$ in R^q : $\theta(z') = 1$ for $|z'| < 1$ and $\theta(z') = 0$ for $|z'| > 1$.

The following assertion holds true.

Theorem 1. Let $\rho(x') \in C_0^{\alpha+1}(\Gamma_1)$. Then near the manifold Γ_1 the solution $u'(x)$ is represented by the following equality

$$\begin{aligned} u'(x) &= I_{11}(x) + \sum_{\lambda=0}^{\alpha-1} Q_\lambda(D') \rho(x') \cdot r_2^{-(n-q-1-\lambda)} \\ &+ a_n (\ln \frac{1}{r_2} + \gamma) Q_\alpha(D') \rho(x') + I_2(x', x_0'') + o(r_2), \\ &\quad o(r_2) \rightarrow 0, r_2 \rightarrow 0 \end{aligned} \quad (11)$$

where γ is a constant, defined in the proof of theorem.

Proof. For establishing the representation (11) we investigate each integral I_1 , I_2 in neighbourhood of the manifold Γ_1 .

1. We begin with the integral I_1 . It is easy to see that

$$\begin{aligned} I_1(x) &= I_{11}(x) + a_n \int_{R^q} r^{-n+2} \cdot P_{\alpha-1}(x', z') \rho dz' + \\ &+ \int_{|z'|<1} \mathcal{P}_\alpha(x', z') \rho dz' = I_{11}(x) + I_{12}(x) + I_{13}(x). \end{aligned} \quad (12)$$

a) Since $\rho(x') \in C_0^{\alpha+1}(R^q)$ the integral $I_{11}(x)$ is absolutely and uniformly convergent on each compact $K = \{x \in R_+^n, |x| < R\}$. Therefore, $I_1(x)$ defines a continuous function near the variety Γ_1 and it is clear that

$$I_{11}(x) = I_{11}(x', x_0'') + o(r_2), \quad o(r_2) = o(|x'' - x_0''|) \rightarrow 0, r_2 \rightarrow 0. \quad (13)$$

b) Now consider $I_{12}(x)$. Firstly, making the substitution $z' = r_2 \cdot \zeta'$, after passing to spherical coordinates (ω', r_1) , we obtain

$$\begin{aligned} I_{12}(x) &= a_n \cdot \sum_{\lambda=0}^{\alpha-1} Q_\lambda(D') \rho(x') \cdot \int_0^\infty r^{2-n} \cdot r_1^{q+\lambda-1} dr_1 \\ &= \sum_{\lambda=0}^{\alpha-1} Q_\lambda(D') \rho(x') \cdot a_n I_\lambda \cdot r_2^{-(n-q-2-\lambda)}, \quad r^2 = r_1^2 + r_2^2. \end{aligned} \quad (14)$$

In I_λ we make the substitution $r_1^2 \cdot (1+r_1^2)^{-1} = v$, then use the relation between Euler's B and Γ functions and the relation $z\Gamma(z) = \Gamma(z+1)$ [10]. Thus $a_n I_\lambda = 2^{-1} \pi^{-n/2} \Gamma((q+\lambda)/2) \cdot \Gamma((\alpha-\lambda)/2) \stackrel{\text{def}}{=} \beta_\lambda$. Therefore,

$$I_{12}(x) = \sum_{\lambda=0}^{\alpha-1} \beta_\lambda Q_\lambda(D') \rho(x') \cdot r_2^{-(\alpha-\lambda)}. \quad (15)$$

c) Now we calculate the integral $I_{13}(x)$.

$$I_{13}(x) = \sum_{|k|=\alpha} \frac{\rho^{(k)}(x')}{k!} a_n \int_{|z'|<1} r^{2-n} \cdot (-z')^k dz' \stackrel{\text{def}}{=} \sum_{|k|=\alpha} a_n B_k \cdot \frac{\rho^{(k)}(x')}{k!}. \quad (16)$$

Firstly, we make the substitution of variables $z' = r_2 \cdot \zeta'$, then pass to spherical coordinate (ω, r_1) and separate the finite part of the integral [8]. We obtain

$$\begin{aligned} B_k &= a_n A_k \int_0^{r_2^{-1}} r_1^{\alpha+q-1} (1+r_1)^{(2-n)/2} dr_1 \\ &= a_n A_k \left[\int_0^1 \xi^{n-2} r_1^{-1} dr_1 + \int_1^{r_2^{-1}} (\xi^{n-2} - 1) \frac{dr_1}{r_1} \right] + a_n A_k \int_1^{r_2^{-1}} \frac{dr_1}{r_1} \\ &= \gamma a_n A_k + a_n A_k \ln \frac{1}{r_2} + o(r_2), \end{aligned} \quad (17)$$

where $\xi^2 = r_1^2(1+r_1^2)^{-1}$, γ stands for the expression from the square brackets in (17).

2. Now consider $I_2(x)$. Evidently, $I_2(x)$ is a continuous function near the variety Γ_1 and, therefore,

$$I_2(x) = I_2(x', x''_0 - \bar{x}_0'') + o(x), \quad o(x) \rightarrow 0, \quad r_2 \rightarrow 0. \quad (18)$$

Placing (13),(14),(15),(16) in (12) and (12),(18) in (9) we obtain the desired representation (11). ■

Now we make the following remarks.

1) Evidently, the representation (11) contains only operators $Q_\lambda(D')\rho(x')$ of even orders;

2) The representation (11) contains the function $\ln 1/r_2$ only when the number $n-q$ is even;

3) The expression $I_{11}(x) = I_{11}(x', x'' - x''_0) = \int_{R^q} E_n(z', 0) [\rho(x' - z') - P_{\alpha-1}(x', z')\rho + \theta(z')\mathcal{P}_\alpha(x', z')\rho] dz'$ is a regularization of the trace of the fundamental solution $E_n(x)$ on the variety $\Gamma_1 = R^q$;

4) The main (singular) member of the representation (11) is the fundamental solution $E_{n-q}(x'')$ of the operator $\Delta_{x''}$ in the orthogonal subspace R^{n-q} ;

5) Establishing representations for the derivatives $\frac{\partial^s u'(x)}{\partial r_2^s}$ near R^q is quite similar.

On the Sobolev problem with singular and regularized boundary conditions on γ_k .

The representation (11) is the sum of the singular part $u'_s(x)$ and the regular part $u'_r(x)$ of the solution $u'(x)$. In the cases when the boundary

conditions (3) on Γ_k lose the sense as traces on Γ_k , it is necessary to change these conditions with other natural boundary conditions. As it was noted before, these conditions are about the asymptotic behaviour of the solution near Γ_k or about giving on Γ_k of a regularized part. Other boundary conditions on Γ_k can be imposed. Here we consider two simple cases of singular boundary conditions on Γ_k .

a) Let G be a bounded domain defined before, Γ_k is smooth q_k -dimensional manifold, locally diffeomorphic to R^{q_k} , $x = (x', x'')$, where $x' \in R^{n_k}$ ($n_k = n - q_k$).

Consider the follows boundary value problem. Find in G the solution of the equation

$$\Delta u(x) = f(x) \quad (19)$$

with usually condition

$$u(x) |_{\Gamma_0} = \varphi_0(x) \quad (20)$$

and singular boundary condition on Γ_k :

$$\lim_{x \rightarrow (x', x''_0) \in \Gamma_k} u(x) E_{n_k}^{-1}(x'' - x''_0) = \rho_k(x'), \quad k = 1, \dots, \chi. \quad (21)$$

The following statement is true

Theorem 2. *Let Γ_k be locally isomorphic to R^{q_k} , $f(x) \in C^0(G)$, $\varphi_0(x) \in C^2(\Gamma_0)$, $\rho_k(x') \in C^0(\Gamma_k)$ and $u(x)$ solution of boundary value problem (19)–(21). Then*

$$\begin{aligned} u(x) = & \int_G \mathcal{G}(x, y) f(y) dy + \int_{\Gamma_0} \frac{\partial \mathcal{G}}{\partial \nu}(x, y) \cdot \varphi_0(y') dy' + \\ & + \sum_{k=1}^{\chi} \int_{\Gamma_k} \mathcal{G}(x, y', x''_0) \rho_k(y') dy', \end{aligned} \quad (22)$$

where $\mathcal{G}(x, y)$ is the Green function of Dirichlet problem (19), (20) in \overline{G} .

Proof is similar to proof of Riemann–Green formulae [9, chap. 2]. Denote by S_k^ε the surface $|x - x_0| = \varepsilon$, where $x_0 \in \Gamma_k$, G_ε is the domain with boundary $\Gamma_\varepsilon = \Gamma_0 \cup \{ \bigcup_{k=1}^{\chi} S_k^\varepsilon \}$, ν_k is the exterior normal to Γ_ε in the point $x \in S_k^\varepsilon$. Writing the Green formulae [9] for function $\mathcal{G}(x, y)$ and $u(x)$ in the domain G_ε , passing to limit when $\varepsilon \rightarrow 0$ and taking in consideration (21), we obtain (22). ■

b) Let $G = R_+^n / R^q$, Γ_0 be the hiperplan $x_n = 0$ in R^n , $\Gamma_1 = R^q$ from the problem (21), (4), (5). Consider the follows boundary problem. Find in G the solution $u(x)$ of the equation (19) with the boundary condition (20) on Γ_0 and the additional boundary condition on the Γ_1 :

$$u_{reg}(x) = h(x'), \quad (23)$$

where $u_{reg}(x)$ is the regular (finite) part of solution $u(x)$. The conditions (23) mean that

$$I_{11}(x') + \gamma a_n Q_\alpha(D') \rho(x') + I_2(x', x_0'') = h(x'), \quad (24)$$

where $h(x')$ is given function.

The equation (24) is a hipersingular integro-differential equation [11], whose type depends on the number $n - q$. Some classes of such equation is investigated in [11].

The solvability of the equation (24) means that BVP with regularized conditions (23) is solvable. Each solution of the equation (24) generated a solution of BVP with regularized BC (23).

References

- [1] SOBOLEV S.L. *Some Applications of Functional Analysis in Mathematical Physics*, Izd. Leningrad. Univ., Leningrad, 1950.
- [2] STERNIN B.YU. *Elliptic and parabolic problems on manifolds whose boundary consist of components of various dimensions*, Trudy Mosk. Mat. Obshch., **15** (1966), 346–382.
- [3] ROITBERG Y.A., SKLYARETS A.V. *Sobolev's problem in complete scales of Banach spaces*. Ukr. Math. J., **48** (1996), no. 11, 1555–1969.
- [4] ROITBERG Y.A., SKLYARETS A.V. *Sobolev's problem in complete scales of Banach spaces*. Dokl. Ukrain. Akad. Nauk, **48** (1996), no. 1.
- [5] ROITBERG Y.A. *Boundary Value Problems in the Spases of Distributions*. Kluwer Acad. Publ. Dordrecht, v. **498**, Boston/London, 1999, -246p.
- [6] NIZHNIK L.P. *On point interaction in quantum mechanics*, Ukr. Mat. Zh., **49** (1997), no. 11, 1557–1560.
- [7] NIZHNIK L.P. *Boundary value problems with singular conditions on boundary components of small dimensions*, Methods Funct. Anal. Topology, vol.7, N1, 2001, pp. 76-81.
- [8] GELFAND I.M. AND SHILOV G.E. *Generalized Functions and Operations over These Functions*, Fizmatgiz, Moscow, 1959.
- [9] BARBU V. *Boundary Value Problems for Partial Differential Equations*, Edit. Acad. Rom. Bucharest, 1993.
- [10] KORN G. AND KORN T. *Mathematical Handbook for Sientists and Engineers*, New York. Toronto. London, 1961.
- [11] SAMCO S.G., KILBAS A.A., MARICHEV O.I. *Integrals and derivatives of fractional order and this applications*, Minsk, "Science and Teknik", 1987.

ON SOME OPTIMIZATION PROBLEM WITH NON-QUADRATIC CRITERION

Adam Kowalewski

Institute of Automatics

University of Mining and Metallurgy

al. Mickiewicza 30, 30-059 Cracow, Poland

ako@ia.agh.edu.pl

Abstract Various optimization problems for linear parabolic systems with multiple constant time delays are considered. In this paper, we consider an optimal distributed control problem for a linear parabolic system in which multiple constant delays appear in the state equation. Sufficient conditions for the existence of a unique solution of the parabolic time delay equation with the Dirichlet boundary condition are proved. The time horizon T is fixed. Making use of the Lions scheme [9], necessary and sufficient conditions of optimality for the Dirichlet problem with non-quadratic criterion and constrained control are derived.

Keywords: distributed control, parabolic systems, multiple constant delays.

Introduction

Various optimization problems associated with the optimal control of distributed parabolic systems with time delays appearing in the boundary conditions have been studied recently in Refs. [1]-[8] and [11],[12].

In this paper, we consider an optimal distributed control problem for a linear parabolic system in which multiple constant time delays appear in the state equation.

Such systems constitute in a linear approximation, a universal mathematical model for many diffusion processes.

Sufficient conditions for the existence of a unique solution of such parabolic equations with the Dirichlet conditions are proved. In this paper, we restrict our considerations to the case of the distributed control for the Dirichlet problem. Consequently, we formulate the following op-

timal control problem. We assume that the performance functional has non-quadratic form. Moreover, the time horizon is fixed in our optimization problem. Finally, we impose some constraints on the distributed control . Making use of the Lions framework [9] necessary and sufficient conditions of optimality for the Dirichlet problem with non-quadratic criterion and constrained control are derived.

1. Existence and uniqueness of solutions

Consider now the distributed-parameter system described by the following parabolic delay equation

$$\frac{\partial y}{\partial t} + A(t)y + \sum_{i=1}^m b_i(x, t)y(x, t - h_i) = v \quad x \in \Omega, t \in (0, T) \quad (1)$$

$$y(x, t') = \Phi_0(x, t') \quad x \in \Omega, t' \in [-h_m, 0) \quad (2)$$

$$y(x, 0) = y_0(x) \quad x \in \Omega \quad (3)$$

$$y(x, t; v)|_{\Sigma} = 0 \quad x \in \Gamma, t \in (0, T) \quad (4)$$

where: $\Omega \subset R^n$ is a bounded, open set with boundary Γ , which is a C^∞ -manifold of dimension $n - 1$. Locally, Ω is totally on one side of Γ ,

$$\begin{aligned} y &\equiv y(x, t; v), \quad v \equiv v(x, t), \\ Q &= \Omega \times (0, T), \quad \bar{Q} = \bar{\Omega} \times [0, T], \\ Q_0 &= \Omega \times [-h_m, 0), \quad \Sigma = \Gamma \times (0, T), \end{aligned}$$

b_i are given real C^∞ functions defined on \bar{Q} ,

h_i are specified positive numbers representing multiple time delays, such that $0 \leq h_1 < h_2 < \dots < h_m$ for $i = 1, \dots, m$,

Φ_0 is an initial function defined on Q_0 .

The operator $A(t)$ has the form

$$A(t)y = - \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left(a_{ij}(x, t) \frac{\partial y(x, t)}{\partial x_j} \right) \quad (5)$$

and the functions $a_{ij}(x, t)$ satisfy the condition

$$\sum_{i,j=1}^n a_{ij}(x, t) \Phi_i \Phi_j \geq \alpha \sum_{i=1}^n \Phi_i^2 \quad \alpha > 0, \quad \forall (x, t) \in \bar{Q}, \quad \forall \Phi_i \in R \quad (6)$$

where: $a_{ij}(x, t)$ are real C^∞ functions defined on \bar{Q} (closure of Q).

The equations (1)–(4) constitute a Dirichlet problem.

First we shall prove sufficient conditions for the existence of a unique solution of the mixed initial-boundary value problem (1)–(4) for the case where $v \in L^q(Q)$, $1 < q < \infty$.

The existence of a unique solution for the mixed initial-boundary value problem (1)–(4) on the cylinder Q can be proved using a constructive method. Using the results of Section 14 ([9], pp. 230-231) we can prove the following result.

Theorem 1.1 *Let y_0, Φ_0 and v be given with $y_0 \in L^q(\Omega)$, $\Phi_0 \in L^q(Q_0)$ and $v \in L^q(Q)$, $1 < q < \infty$. Then, there exists a unique solution $y \in L^q(Q)$ for the mixed initial-boundary value problem (1)–(4). Moreover, $\frac{\partial y(v)}{\partial x_i} \in L^q(Q)$, $i = 1, \dots, n$.*

Remark 1.1 We refer to [10] for the solution of parabolic problems in L^q , $q \neq 2$.

2. Problem formulation. Optimization theorem

We shall now formulate the optimal control problem for the Dirichlet problem. Let us denote by $U = L^q(Q)$ the space of controls. The time horizon T is fixed in our problem.

The performance functional is given by

$$I(v) = \lambda_1 \int_Q |y(x, t; v) - z_d|^q dx dt + \lambda_2 \|v\|_{L^q(Q)}^q \quad (7)$$

where: $\lambda_i \geq 0$, $\lambda_1 + \lambda_2 > 0$; z_d is a given element in $L^q(Q)$.

Finally, we assume the following constraint on controls $v \in U_{ad}$, where

$$U_{ad} \text{ is a closed, convex subset of } U \quad (8)$$

Let $y(x, t; v)$ denote the solution of the mixed initial-boundary value problem (1)–(4) at (x, t) corresponding to a given control $v \in U_{ad}$. We note from the Theorem 1.1 that for any $v \in U_{ad}$ the performance functional (7) is well-defined since $y(v) \in L^q(Q)$. The solving of the formulated optimal control problem is equivalent to seeking a $v_0 \in U_{ad}$ such that $I(v_0) \leq I(v) \quad \forall v \in U_{ad}$.

Remark 2.1 If $q \neq 2$, U is a reflexive Banach space (but not Hilbert); but as indicated in [9] (p. 8), results of the type of Theorem 1.3 ([9], p.10) remain valid.

Then from the Theorem 1.3 ([9], p. 10) - in the case where U is a reflexive Banach space - it follows that for $\lambda_2 > 0$ a unique optimal control v_0 exists; moreover, v_0 is characterized by the following condition

$$I'(v_0) \cdot (v - v_0) \geq 0 \quad \forall v \in U_{ad} \quad (9)$$

But from (7), we have for $w \in U$

$$\left. \begin{aligned} I'(v_0) w &= \frac{d}{d\theta} I(v_0 + \theta w) \Big|_{\theta=0} = \\ &= \lambda_1 q \int_Q |y(v_0) - z_d|^{q-2} (y(v_0) - z_d) \frac{\partial}{\partial v_0} y(v_0) w \, dxdt + \\ &\quad + q \lambda_2 \int_Q |v_0|^{q-2} v_0 w \, dxdt \end{aligned} \right\} \quad (10)$$

where

$$\frac{\partial}{\partial v_0} y(v_0) w = \Psi(w) \quad (11)$$

Then we may verify that

$$\frac{\partial \Psi(w)}{\partial t} + A\Psi(w) + \sum_{i=1}^m b_i(x, t)\Psi(x, t-h_i; w) = w, \quad x \in \Omega, \quad t \in (0, T) \quad (12)$$

$$\Psi(x, t') = 0 \quad x \in \Omega, \quad t' \in [-h_m, 0) \quad (13)$$

$$\Psi(x, 0; w) = 0 \quad x \in \Omega \quad (14)$$

$$\Psi(w)|_{\Sigma} = 0 \quad x \in \Gamma, \quad t \in (0, T) \quad (15)$$

Consequently, we can express (9) (dividing through by q) in the following form

$$\begin{aligned} &\lambda_1 \int_Q |y(v_0) - z_d|^{q-2} (y(v_0) - z_d) \Psi(v - v_0) dxdt + \\ &+ \lambda_2 \int_Q |v_0|^{q-2} v_0 (v - v_0) dxdt \geq 0 \quad \forall v \in U_{ad} \end{aligned} \quad (16)$$

To simplify (16), we introduce the adjoint equation and for every $v \in U_{ad}$, we define the adjoint variable $p = p(v) = p(x, t; v)$ as the solution of the equation

$$\begin{aligned} &-\frac{\partial p(v)}{\partial t} + A^*(t)p(v) + \sum_{i=1}^m b_i(x, t+h_i)p(x, t+h_i; v) = \\ &= \lambda_1 |y(v) - z_d|^{q-2} (y(v) - z_d), \quad x \in \Omega, \quad t \in (0, T - h_m) \end{aligned} \quad (17)$$

$$-\frac{\partial p(v)}{\partial t} + A^*(t)p(v) = \lambda_1 |y(v) - z_d|^{q-2} (y(v) - z_d), \quad x \in \Omega, \quad t \in (T - h_m, T) \quad (18)$$

$$p(x, T; v) = 0 \quad x \in \Omega \quad (19)$$

$$p(x, t; v) = 0 \quad x \in \Gamma, \quad t \in (0, T) \quad (20)$$

$$\text{where } A^*(t)p = -\sum_{i,j=1}^n \frac{\partial}{\partial x_j} \left(a_{ij}(x, t) \frac{\partial p}{\partial x_i} \right) \quad (21)$$

The existence of a unique solution for the problem (17)–(20) on the cylinder Q can be proved using a constructive method. It is easy to notice that for given z_d and v , problem (17)–(20) can be solved backwards in

time starting from $t = T$, i.e., first, solving (17)–(20) on the subcylinder Q_K and in turn on Q_{K-1} , etc. until the procedure covers the whole cylinder Q . For this purpose, we may apply Theorem 1.1 (with an obvious change of variables) to problem (17)–(20) (with reversed sense of time, i.e., $t' = T - t$).

Lemma 2.1 *Let the hypothesis of Theorem 1.1 be satisfied. Then, for given $z_d \in L^q(Q)$ and any $v \in L^q(Q)$, there exists a unique solution $p(v) \in L^q(Q)$ for the problem (17)–(20). Moreover,*

$$\frac{\partial p(v)}{\partial x_i}, \frac{\partial p(v)}{\partial t}, \frac{\partial^2}{\partial x_i \partial x_j} p(v) \in L^{q'}(Q)$$

and

$$|y(v_0) - z_d|^{q-2} (y(v_0) - z_d) \in L^{q'}(Q)$$

$$\frac{1}{q} + \frac{1}{q'} = 1.$$

We simplify (16) using the adjoint equation (17)–(20). For this purpose setting $v = v_0$ in (17)–(20), multiplying both sides of (17), (18) by $\Psi(v - v_0)$, then integrating over $\Omega \times (0, T - h_m)$ and $\Omega \times (T - h_m, T)$ respectively and then adding both sides of (17), (18) we get

$$\begin{aligned} \lambda_1 \int_Q |y(v_0) - z_d|^{q-2} (y(v_0) - z_d) \Psi(v - v_0) dx dt &= \\ &= \int_Q \left(-\frac{\partial p(v_0)}{\partial t} + A^*(t)p(v_0) \right) \Psi(v - v_0) dx dt + \\ &+ \sum_{i=1}^m \int_0^{T-h_m} \int_\Omega b_i(x, t + h_i) p(x, t + h_i; v_0) \Psi(v - v_0) dx dt = \quad (22) \\ &= \int_Q p(v_0) \frac{\partial}{\partial t} \Psi(v - v_0) dx dt + \int_Q A^*(t)p(v_0) \Psi(v - v_0) dx dt + \\ &+ \sum_{i=1}^m \int_0^{T-h_m} \int_\Omega b_i(x, t + h_i) p(x, t + h_i; v_0) \Psi(v - v_0) dx dt \end{aligned}$$

Using equation (1), the first integral on the right-hand side of (22) can be rewritten as

$$\begin{aligned}
\int_Q p(v_0) \frac{\partial}{\partial t} \Psi(v - v_0) dxdt &= - \int_Q p(v_0) A(t) \Psi(v - v_0) dxdt - \\
&- \sum_{i=1}^m \int_0^T \int p(x, t; v_0) b_i(x, t) \Psi(x, t - h_i; v - v_0) dxdt + \\
&+ \int_Q p(v_0) (v - v_0) dxdt = - \int_Q p(v_0) A(t) \Psi(v - v_0) dxdt - \\
&- \sum_{i=1}^m \int_{-h_i}^{T-h_i} \int_{\Omega} p(x, t_i + h_i; v_0) b_i(x, t_i + h_i) \Psi(x, t_i; v - v_0) dxdt_i + \\
&+ \int_Q p(v_0) (v - v_0) dxdt
\end{aligned} \tag{23}$$

where: $t_i = t - h_i$ and $dt = dt_i$.

Substituting (23) into (22), after transformations we obtain

$$\begin{aligned}
\lambda_1 \int_Q |y(v_0) - z_d|^{q-2} (y(v_0) - z_d) \Psi(v - v_0) dxdt &= \\
&= \int_Q p(v_0) (v - v_0) dxdt
\end{aligned} \tag{24}$$

Substituting (24) into (16) we obtain

$$\int_Q (p(v_0) + \lambda_2 |v_0|^{q-2} v_0) (v - v_0) dxdt \geq 0, \quad \forall v \in U_{ad} \tag{25}$$

Theorem 2.1 *For the problem (1)–(4) with the performance functional (7) with $z_d \in L^q(Q)$ and $\lambda_2 > 0$ and with constraints on controls (8), there exists a unique optimal control v_0 which satisfies the maximum condition (25).*

Consider now the particular case where $U_{ad} = L^q(Q)$.

Thus the maximum condition (25) is satisfied when

$$v_0 = -\frac{1}{\lambda_2^{q'-1}} |p|^{q'-2} p \tag{26}$$

We must notice that the conditions of optimality derived above (Theorem 2.1) allow us to obtain an analytical formula for the optimal control in particular cases only (e.g. there are no constraints on controls). This results from the following: the determining of the function $p(v_0)$ in the maximum condition (25) is possible from the adjoint equation (17)–(20) if and only if we know y_0 which corresponds to the control v_0 . These mutual connections make the practical use of the derived optimization formulas difficult. Therefore we resign from the exact determining of the optimal control and we use approximation methods.

Remark 2.2 We can also consider a non-quadratic form on a Hilbert space: take $U = L^2(Q)$; then $y(v) \in L^q(Q)$ if $1 < q < 2$ and hence, in particular, we can consider the performance functional (7) with $1 < q < 2$.

3. Conclusions

The results presented in the paper can be treated as a generalization of the results obtained in [9] onto the case of multiple constant time delays appearing in the state equations.

We can also obtain estimates and a sufficient condition for the boundedness of solutions for such parabolic time delay systems.

Finally, we can consider optimal control problems of hyperbolic systems with multiple time delays appearing in the state equations.

The ideas mentioned above will be developed in forthcoming papers.

Acknowledgments

The research presented here was carried out within the research programmes University of Mining and Metallurgy, No. 10.10.120.31 and No. 10.10.120.40.

References

- [1] Knowles, G. (1978). Time-optimal control of parabolic systems with boundary conditions involving time delays. *Journal of Optimization Theory and Applications*, 25:563–574.
- [2] Kowalewski, A. (1987). Optimal control with initial state not a priori given and boundary condition involving a delay. *Lecture Notes in Control and Information Sciences*, Berlin, Heidelberg: Springer-Verlag, 95:94–108,
- [3] Kowalewski, A. (1988). Boundary control of distributed parabolic system with boundary condition involving a time-varying lag. *International Journal of Control*, 48:2233–2248.
- [4] Kowalewski, A. (1990). Feedback control for a distributed parabolic system with boundary condition involving a time-varying lag. *IMA Journal of Mathematical Control and Information*, 7:143–157.
- [5] Kowalewski, A. (1990). Minimum time problem for a distributed parabolic system with boundary condition involving a time-varying lag. *Archives of Automatics and Remote Control*, XXXV, 3-4:145–153.
- [6] Kowalewski, A. (1990). Optimality conditions for a parabolic time delay system. *Lecture Notes in Control and Information Sciences*, Berlin, Heidelberg: Springer-Verlag, 144:174–183.
- [7] Kowalewski, A. (1993). Optimal control of parabolic systems with time-varying lags. *IMA Journal of Mathematical Control and Information*, 10:113–129.

- [8] Kowalewski, A. and Duda, J. (1992). On some optimal control problem for a parabolic system with boundary condition involving a time-varying lag. *IMA Journal of Mathematical Control and Information*, 9:131–146.
- [9] Lions, J.L. (1971). *Optimal Control of Systems Governed by Partial Differential Equations*. Berlin, Heidelberg: Springer-Verlag.
- [10] Solonnikov, V.A. (1965). Mixed problem for a Linear Parabolic Equations in L^p . *Papers of Steklov Mathematical Institute*, 83:3–162 (in Russian).
- [11] Wang, P.K.C. (1975). Optimal control of parabolic systems wih boundary conditions involving time delays. *SIAM Journal of Control*, 13:274–293.
- [12] Wong, K.H.: (1987). Optimal control computation for parabolic systems with boundary conditions involving time delays. *Journal of Optimization Theory and Applications*, 53:475–507.

NONCONSERVATIVE SCHRÖDINGER EQUATIONS WITH UNOBSERVED NEUMANN B.C.: GLOBAL UNIQUENESS AND OBSERVABILITY IN ONE SHOT*

I. Lasiecka and R. Triggiani

Department of Mathematics, University of Virginia, Charlottesville, VA 22904 USA
il2v@weyl.math.virginia.edu, rt7u@virginia.edu

X. Zhang

Department of Mathematics, Sichuan University, Chengdu 610064 PRC
Departamento de Matematica Aplicada, Universidad Complutense, 28040 Madrid, Spain

1. Introduction. Problem statement

1.1. Problem statement. Assumptions

Let Ω be an open bounded domain in \mathbb{R}^n with boundary $\partial\Omega = \Gamma$ of class C^2 , consisting of the closure of two disjoint parts: Γ_0 (uncontrolled or unobserved part) and Γ_1 (controlled or observed part), both relatively open in Γ : $\partial\Omega = \Gamma \equiv \overline{\Gamma}_0 \cup \overline{\Gamma}_1$, $\Gamma_0 \cap \Gamma_1 = \emptyset$. In this paper, we consider the following Schrödinger equation in the (complex-valued) unknown $w(t, x)$ defined on Ω :

$$\mathcal{P}w \equiv iw_t + \Delta w = F(w) + f \quad \text{in } Q \equiv (0, T] \times \Omega. \quad (1.1.1)$$

In (1.1.1), we have set

$$F(w) \equiv q_1(t, x) \cdot \nabla w(t, x) + q_0(t, x)w(t, x), \quad (1.1.2)$$

subject to the following preliminary *standing assumption* on the coefficients: We let $|q_1|, q_0 \in L_\infty(Q)$, so that the following pointwise estimate

*The work of I.Lasiecka and R.Triggiani was partially supported by the National Science Foundation under Grant DMS-0104305 and by the Army Research Office under Grant DAAD19-02-1-0179. The work of X. Zhang was partially supported by the Foundation for the Author of National Excellent Doctoral Dissertation of China (project No. 200119); Grant BFM2002-033450 of the Spanish MCYT; and the NSF of China under Grant 19901024.

holds true:

$$|F(w)|^2 \leq C_T[|\nabla w|^2 + |w|^2], \quad \forall (t, x) \in Q. \quad (1.1.3)$$

We assume throughout that the non-homogeneous term f satisfies

$$f \in L_2(Q). \quad (1.1.4)$$

This is an announcement of results in the full-length paper [8].

Main assumptions. The main focus of the present paper refers to the case where Eqn. (1.1.1) is supplemented by *purely Neumann* B.C.: $\frac{\partial w}{\partial \nu}|_{\Sigma} \equiv 0$, while observation or control takes place only on a *subportion* Γ_1 of the boundary Γ . This problem is of interest on both physical grounds (it arises in the structural acoustic problem) and on mathematical grounds (the Lopatinski's condition is not satisfied). For this key case, we shall need the following assumptions just as in [7] in the corresponding wave equation case. In addition to the standing hypotheses (1.1.3) on $F(w)$ and (1.1.4) on f , the following assumptions are postulated throughout.

(A.1) Given the triple $\{\Omega, \Gamma_0, \Gamma_1\}$, $\partial\Omega = \overline{\Gamma_0 \cup \Gamma_1}$, there exists a strictly convex (real-valued) non-negative function $d : \overline{\Omega} \rightarrow \mathbb{R}^+$, of class $C^3(\overline{\Omega})$, such that, if we introduce the (conservative) vector field $h(x) \equiv \nabla d(x)$, $x \in \Omega$, then the following two properties hold true:

(i)

$$\left. \frac{\partial d}{\partial \nu} \right|_{\Gamma_0} = \nabla d \cdot \nu = h \cdot \nu = 0 \text{ on } \Gamma_0; \quad h \equiv \nabla d. \quad (1.1.5_N)$$

(ii) the (symmetric) Hessian matrix \mathcal{H}_d of $d(x)$ [i.e., the Jacobian matrix J_h of $h(x)$] is strictly positive definite on $\overline{\Omega}$: there exists a constant $\rho > 0$ such that for all $x \in \overline{\Omega}$:

$$\mathcal{H}_d(x) = J_h(x) = [d_{x_i x_j}] = \left[\frac{\partial h_i}{\partial x_j} \right] \geq \rho I, \quad i, j = 1, \dots, n. \quad (1.1.6)$$

(A.2) A working assumption, which can be relaxed (see Remark 1.1.1), is that $d(x)$ has no critical point on $\overline{\Omega}$:

$$\inf_{x \in \Omega} |h(x)| = \inf_{x \in \Omega} |\nabla d(x)| = p > 0. \quad (1.1.7)$$

Remark 1.1.1. Assumption (A.2) can, in fact, be removed as in [7, Sect. 10], [19, Sec. 10], [8, Sect. 10], by splitting Ω as $\Omega = \Omega_1 \cup \Omega_2$, for two suitable overlapping sets Ω_1 and Ω_2 , and working with two strictly convex functions d_1 and d_2 . Since the full statement of results without

(A.2) requires a rather lengthy preparatory background, we shall here retain assumption (A.2) and refer to the above references for its removal.

Remark 1.1.2. (Neumann) Assumption (1.1.5) is due to the purely Neumann B.C. $\frac{\partial w}{\partial \nu}|_{\Sigma} \equiv 0$, as in the case of the wave equation [7, Remark 1.1.2]. It was introduced in [16, Section 5]. Reference [7, Appendices A-C] provides, by different mathematical techniques, *several classes of triples $\{\Omega, \Gamma_0, \Gamma_1\}$ in \mathbb{R}^n , $n \geq 2$, where assumptions (A.1) and (A.2) are satisfied*. (In light of Remark 1.1.1, only assumption (A.1) is the critical one.) For instance, just to quote one general result: one can construct *explicitly* [7, Theorem A.4.1, p. 301] the required strictly convex function $d(x)$ satisfying (A.1) if the (Euclidean) bounded domain $\Omega \in \mathbb{R}^n$ is (i) convex (respectively, concave) on the side of the portion Γ_0 of its boundary, and (ii) there exists a radial vector field $(x - x_0)$ for some $x_0 \in \mathbb{R}^n$ which is entering (respectively, exiting) Ω through Γ_0 . Moreover, [7, Corollary A.4.2, p. 306] the previously constructed strictly convex function $d(x)$ has the following additional property: its gradient $\nabla d|_{\Gamma_0}$, once restricted on the portion Γ_0 of the boundary, vanishes at the unique point $x \in \Gamma_0$, if such exists on Γ_0 , where the vector field $x - x_0$ is orthogonal to Γ_0 . The above condition on the existence of such $d(x)$ is only sufficient. It holds true also in the case where Ω is a bounded set of a finite-dimensional Riemann manifold [19, Appendix B]. Other classes are given in [7, Appendices A-C] satisfying assumption (A.1): for instance, where the portion Γ_0 of the boundary is logarithmic concave [7, Lemma A.2.2, p. 294].

(Dirichlet) Even though the purely Neumann B.C. case: $\frac{\partial w}{\partial \nu}|_{\Sigma} \equiv 0$ will be the central focus of this paper, our treatment will allow us to also include the purely Dirichlet B.C. case: $w|_{\Sigma} \equiv 0$. Here, however, hypothesis (1.1.5_N) can be dispensed with. It will be replaced by the much weaker condition

$$h \cdot \nu \leq 0 \quad \text{on } \Gamma_0. \tag{1.1.5D}$$

A specific example is $d(x) = \frac{1}{2}\|x - x_0\|^2$, with x_0 outside Ω , where then $h(x) = \nabla d(x) = (x - x_0)$ is radial. [A combination of Dirichlet/Neumann B.C. is also included in our treatment.]

Remark 1.1.3. We expect the techniques and results of the present note (paper [8]) to be extended to the fully general Euclidean case, where the Euclidean Laplacian operator Δ in (1.1.1) is replaced by a second-order elliptic operator with *variable coefficients* (in space), of class, say, $C^2(\overline{\Omega})$; or, more generally, by the Laplace-Beltrami operator Δ_g , in case Ω is a bounded set of a finite-dimensional Riemann manifold $\{M, g\}$, with metric g . This extension would be accomplished, in the present

Schrödinger case, in the same way as the Euclidean wave equation paper [7] (respectively, [4]) was extended to the Riemann wave equation in [19] (no lower-order terms in the estimates) (respectively, [5], with lower-order terms in the estimates); or the Euclidean Schrödinger equation paper [17] was extended to the Riemann Schrödinger equation in [18] (with lower-order terms in the estimates). That is, by replacing the Euclidean metric of the present paper with the appropriate Riemann metric.

At any rate, throughout Section 2.1, we shall merely deal with smooth solutions of Eqn. (1.1.1) with no B.C. imposed, subject *only* to hypotheses (1.1.6) and (1.1.7). Then, assumption (1.1.5_N): $h \cdot \nu \equiv 0$ on Γ_0 [resp. (1.1.5_D): $h \cdot \nu \leq 0$ on Γ_0] will be introduced only when analyzing purely Neumann B.C. [resp. purely Dirichlet B.C.].

Pseudo-convex function $\varphi(x, t)$. Having chosen, by assumption (A.1), a strictly convex potential function $d(x) \geq 0$, we next introduce the pseudo-convex function $\varphi : \Omega \times \mathbb{R} \rightarrow \mathbb{R}$ of class C^3 :

$$\varphi(x, t) = d(x) - c \left(t - \frac{T}{2} \right)^2; \quad 0 \leq t \leq T, \quad x \in \Omega, \quad (1.1.8a)$$

where $T > 0$ is arbitrary, and where then $c = c_T$ is chosen large enough as to have

$$cT^2 > 4 \max_{x \in \bar{\Omega}} d(x), \text{ so that } cT^2 > 4 \max_{x \in \bar{\Omega}} d(x) + 4\delta \quad (1.1.8b)$$

for a suitably small $\delta > 0$, henceforth kept fixed. Unless otherwise explicitly noted, $\varphi(x, t)$ is selected as described above and kept fixed henceforth. Such function $\varphi(x, t)$ has the following properties:

- (a) for the constant $\delta > 0$, fixed in (1.1.8b), we have

$$\varphi(x, 0) \equiv \varphi(x, T) = d(x) - c \frac{T^2}{4} \leq -\delta, \text{ uniformly in } x \in \Omega; \quad (1.1.9)$$

- (b) there are t_0 and t_1 , with $0 < t_0 < \frac{T}{2} < t_1 < T$, such that

$$\min_{x \in \bar{\Omega}, t \in [t_0, t_1]} \varphi(x, t) \geq -\frac{\delta}{2}, \quad (1.1.10)$$

since $\varphi(x, \frac{T}{2}) = d(x) \geq 0$ for all $x \in \Omega$ (in fact, only the weaker property: $\min \varphi(x, t) \geq \sigma > -\delta$ is actually needed).

Throughout this paper, we set

$$E(t) \equiv \int_{\Omega} |\nabla w(t)|^2 d\Omega;$$

$$\mathbb{E}(t) = \int_{\Omega} [|\nabla w(t)|^2 + |w(t)|^2] d\Omega = \|w(t)\|_{H^1(\Omega)}^2. \quad (1.1.11)$$

Goals. As already mentioned, we consider, at first, sufficiently smooth solutions $w(t, x)$ of the Schrödinger Eqn. (1.1.1).

First goal. Then, our *first* goal is to establish Carleman-type inequalities with w in $H^1(\Omega)$ —basic energy level—for these solutions *without lower-order terms*. Carleman inequalities, however, with (interior) lower-order terms were obtained in [17, Theorems 2.1.1 and 2.1.2, pp. 464–466] for the Schrödinger Eqn. (1.1.1), and in [18, Theorems 3.3 and 3.4, pp. 640–641] for Eqn. (1.1.1) with the (Euclidean) Laplacian Δ replaced by a variable coefficient (in space) uniformly elliptic operator; or, with essentially the same effort, for the Schrödinger Eqn. (1.1.1) with the Euclidean Laplacian Δ replaced by the Laplace-Beltrami operator Δ_g , defined on a bounded set $\Omega \subset M$ with boundary, of a Riemann manifold $\{M, g\}$. In the present work, as well as in the prior references [17] and [18], the boundary terms (traces of w) of the solutions w of Eqn. (1.1.1) which appear in the Carleman estimates are given *explicitly*.

Second goal. As a consequence of Carleman estimates without lower-order term and explicit boundary terms (our first goal), we then achieve our second goal: that is, we obtain global uniqueness results as well as continuous observability/uniform stabilization inequalities, with w in the basic energy level $H^1(\Omega)$, in one shot, as part of the same flow of arguments. This is in contrast with prior Carleman estimates, hence continuous observability/uniform stabilization inequalities polluted by lower-order terms as in [12], [18] and, for $F(w) \equiv 0$, in [3]. In this latter case, a first disadvantage is the necessity to require an independent global uniqueness result in order to absorb, and hence eliminate, the lower-order term from the sought-after estimates. Moreover, a second disadvantage is the lack of control on the constants arising in the final estimates, as the aforementioned absorption process proceeds by contradiction.

In short, the key aim of the present paper is to eliminate lower-order terms from the sought-after continuous observability/uniform stabilization estimates, therefore avoiding the two disadvantages cited above.

Third goal. Our third goal is to obtain lower-level energy estimate (more precisely, w in $L_2(\Omega)$ rather than in $H^1(\Omega)$), such as they are needed in the problem of uniform stabilization with $L_2(0, \infty; L_2(\Gamma))$ -Dirichlet feedback control. See Section 2.5.

Literature. As to the literature, we note that prior works on Schrödinger equations [3],[11],[10],[12]–[13],[17],[1],[18] under various B.C. and degree of generality obtain the continuous observability/uniform stabi-

zation inequalities *polluted by lower-order terms*. In fact this is the case for almost all papers on exact controllability/uniform stabilization of the literature, regardless of the specific evolution equation. To our knowledge, the only exceptions where lower-order terms do not appear in the observability/stabilization estimates are [2],[7] for second-order hyperbolic equations with Δ as principal part; [19] for Δ_g as principal part; and [15] for general evolution equations mostly in the case of Dirichlet B.C. The present paper ([8]) – and its successor with Δ replaced by Δ_g – emphasizes the purely Neumann B.C. case of (1.1.1), as is the case for [7].

2. Main results under (A.1) and (A.2)

2.1. Carleman estimates without lower-order terms for $H^{2,2}(Q)$ -solutions

Theorem 2.1.1. (First version) *Let $T > 0$ arbitrary and let $c = c_T$ be defined by (1.1.8b). Let $d(x) \in C^3(\bar{\Omega})$ be the non-negative, real, strictly convex function satisfying assumptions (A.1(ii)) = (1.1.6) and (A.2) = (1.1.7). Define $\varphi(x, t)$ by (1.1.8). Let w be a solution of Eqn. (1.1.1) [with no boundary conditions imposed] in the class:*

$$w \in H^{2,2}(Q) \equiv L_2(0, T; H^2(\Omega)) \cap H^2(0, T; L_2(\Omega)), \quad (2.1.1)$$

where (1.1.1) is subject to the standing assumption (1.1.3) for $F(w)$ and (1.1.4) for f . Then, for all τ sufficiently large, the following one-parameter family of estimates holds true:

$$\begin{aligned} B_\Sigma(w) + 4 \int_0^T \int_{\Omega} e^{2\tau\varphi} |f|^2 d\Omega dt \\ \geq \left[4\tau\rho - \frac{1}{2} - 4C_T \right] \int_0^T \int_{\Omega} e^{2\tau\varphi} |\nabla w|^2 d\Omega dt \\ + [4\tau^3\rho p + \mathcal{O}(\tau^2) - 4C_T] \int_0^T \int_{\Omega} e^{2\tau\varphi} |w|^2 d\Omega dt \\ - c_{d,T}\tau e^{-2\tau\delta} [\mathbb{E}(T) + \mathbb{E}(0)] \\ \geq [k_\tau^2] e^{-\delta\tau} \int_{t_0}^{t_1} \mathbb{E}(t) dt - c_\varphi \tau e^{-2\tau\delta} [\mathbb{E}(T) + \mathbb{E}(0)], \end{aligned} \quad (2.1.2)$$

where $\rho > 0$, $p > 0$, $\delta > 0$ are defined by (1.1.6), (1.1.7), (1.1.8b) and k_τ^2 is a positive constant $\geq k_{\tau_0}^2$, $\forall \tau \geq \tau_0 > 0$. Moreover, $\mathbb{E}(t)$ is defined by (1.1.11), while t_0, t_1 are as in (1.1.10). Finally, setting $h = \nabla d$ as in (1.1.5_N), then the boundary terms $B_\Sigma(w)$ are given explicitly as follows,

where $\xi = \operatorname{Re} w$, $\eta = \operatorname{Im} w$:

$$\begin{aligned}
 B_\Sigma(w) &= 2\tau \int_0^T \int_\Gamma e^{2\tau\varphi} [2\tau^2|h|^2] |w|^2 h \cdot \nu d\Gamma dt \\
 &\quad - 4c\tau \int_0^T \int_\Gamma e^{2\tau\varphi} \left(t - \frac{T}{2} \right) \left[\eta \frac{\partial \xi}{\partial \nu} - \xi \frac{\partial \eta}{\partial \nu} \right] d\Gamma dt \\
 &\quad - 2\tau \int_0^T \int_\Gamma e^{2\tau\varphi} [\xi_t \eta - \xi \eta_t] h \cdot \nu d\Gamma dt \\
 &\quad + \int_0^T \int_\Gamma e^{2\tau\varphi} [2\tau^2|h|^2 - \tau \Delta d] \left[\bar{w} \frac{\partial w}{\partial \nu} + w \frac{\partial \bar{w}}{\partial \nu} \right] d\Gamma dt \\
 &\quad + 2\tau \int_0^T \int_\Gamma e^{2\tau\varphi} h \cdot \left[\nabla \bar{w} \frac{\partial w}{\partial \nu} + \nabla w \frac{\partial \bar{w}}{\partial \nu} \right] d\Gamma dt \\
 &\quad - 2\tau \int_0^T \int_\Gamma e^{2\tau\varphi} |\nabla w|^2 h \cdot \nu d\Gamma dt.
 \end{aligned} \tag{2.1.3a}$$

$$\begin{aligned}
 B_\Sigma(w) &\leq C_\tau \left\{ \int_0^T \int_\Gamma e^{2\tau\varphi} \left[|w|^2 + \left| \frac{\partial w}{\partial \nu} \right|^2 \right] d\Sigma \right. \\
 &\quad \left. + \int_0^T \left[\|w_t\|_{H^{-1}(\Gamma)}^2 + \|w\|_{H^1(\Gamma)}^2 \right] dt \right\}.
 \end{aligned} \tag{2.1.3b}$$

Theorem 2.1.1 is proved in [8, Sections 3 through 5].

Remark 2.1.1. The Carleman estimate (2.1.2) of Theorem 2.1.1 is essentially the one in [17, Theorem 2.1.1, p. 464] later generalized in [18, Theorem 3.3, p. 640] to the case where the Euclidean Laplacian Δ is replaced by the Laplace-Beltrami operator Δ_g on a Riemann manifold $\{M, g\}$, except for the critical improvement that our present version (2.1.2) does not include an interior lower-order term, unlike the estimates of the aforementioned references [17], [18].

As in [17], [18], in order to refine Theorem 2.1.1, we need to specialize the first-order differential operator $F(w)$ by imposing a *structural* property, as stated by the following assumption:

(A.3) Let $F(w)$ in (1.1.2) be specialized as follows:

$$\begin{cases} q_1(t, x) = \nabla \pi(t, x) - ir_1(t, x), \text{ for a real-valued scalar function} \\ \pi(t, x) \text{ and a real-valued vector function } r_1(t, x), \\ r_1 \in L_\infty(0, T; [H^1(\Omega)]^n), \quad q_0 \in L_\infty(0, \infty; H^1(\Omega)). \end{cases} \tag{2.1.4}$$

[At the price of a change of variable, one may take $\pi(t, x) \equiv 0$. Then, the structural property $iq_1 = r_1$ (real) is *critical*, but we are not attempting to extract the minimal possible regularity of r_1 and q_0 .]

The use of assumption (A.3) = (2.1.4) is seen in [8, Lemma 6.1(ii),(iv), Eqn. (6.7), (6.9), estimate (6.12) and Remark 6.1]. Such assumption, which is related to well-posedness, will permit us to obtain a second version, more refined, of the Carleman estimate (2.1.2) of Theorem 2.1.1.

Theorem 2.1.3. (Second version) *Assume the setting (in particular, (1.1.6), (1.1.7)), and the notation of Theorem 2.1.1. In addition, we assume hypothesis (A.3) = (2.1.4) and that $f \in L_2(0, T; H^1(\Omega))$. Let w be a solution of Eqn. (1.1.1) in the class $H^{2,2}(Q)$ in (2.1.1). Then, for all $\tau > 0$ sufficiently large, the following one-parameter family of estimates holds true, where $k_{\varphi,\tau} > 0$,*

$$\begin{aligned} \tilde{B}_\Sigma(w) + 4 \int_0^T \int_\Omega e^{2\tau\varphi} |f|^2 d\Omega dt + C_{p,q,\rho,\tau} \|f\|_{L_2(0,T;H^1(\Omega))}^2 \\ \geq \{[4\tau\rho - \frac{1}{2} - 4C_T]e^{-\delta\tau} \frac{(t_1-t_0)}{2} e^{-C_T T} - c_{d,T}\tau e^{-2\tau\delta}\} [\mathbb{E}(T) + \mathbb{E}(0)] \end{aligned} \quad (2.1.5)$$

$$\geq k_{\varphi,\tau} [\mathbb{E}(T) + \mathbb{E}(0)] \geq k_{\varphi,\tau} [E(T) + E(0)], \quad (2.1.6)$$

with explicit constant which is noted in [8, Eqn. (6.21)]. $\mathbb{E}(t)$ and $E(t)$ are defined in (1.1.11). Moreover, the boundary terms $B_\Sigma(w)$ in (2.1.5) are given by

$$\begin{aligned} \tilde{B}_\Sigma(w) &= B_\Sigma(w) + C_{\varphi,p,q} \left\{ \int_0^T \int_\Gamma \left| \frac{\partial w}{\partial \nu} \right| \left[|w_t| + \left| \frac{\partial w}{\partial \mu} \right| |r_1 \cdot \mu| \right. \right. \\ &\quad + \left. \left| \frac{\partial w}{\partial \nu} \right| |r_1 \cdot \nu| + |q_0 w| + |w| + |f| \right] d\Gamma dt \\ &\quad \left. + \int_0^T \int_\Gamma |\nabla w|^2 |r_1 \cdot \nu| d\Gamma dt \right\} + \int_0^T \int_{\Gamma_1} |w|^2 d\Gamma_1 dt, \end{aligned} \quad (2.1.7)$$

with constant $C_{\varphi,p,q}$ given explicitly in the left side of [8, (6.35)], where $\frac{\partial}{\partial \mu}$ denotes the tangential derivative, μ is a unit tangent vector. [Moreover, the last boundary term $\int_0^T \int_{\Gamma_1}$ in (2.1.7) may be omitted, if it is known that w vanishes on a portion of the boundary of positive measure, so that Poincaré inequality holds true.]

Theorem 2.1.2 is proved in [8, Section 6] (for $H^{2,2}(Q)$ -solutions).

2.2. Extension of Carleman estimates to finite energy solutions

Theorem 2.2.1. *Assume the hypotheses of Theorem 2.1.2: (1.1.3), (A.3) = (2.1.4) for $F(w)$, $f \in L_2(0, T; H^1(\Omega))$, in addition to (A.1(ii)) = (1.1.6) and (A.2) = (1.1.7). Then, estimate (2.1.2) of Theorem 2.1.1*

and estimate (2.1.6) of Theorem 2.1.2 can be extended to finite energy solutions in the class

$$w \in C([0, T]; H^1(\Omega)); \frac{\partial w}{\partial \nu} \in L_2(0, T; L_2(\Gamma)), w_t \in L_2(0, T; H^{-1}(\Gamma)), \quad (2.2.1)$$

provided that the (unbounded) term $F(w) = -ir_1 \cdot \nabla w + q_0 w$ satisfies the following additional hypothesis:

(A.4) (this is only for convenience)

the coefficients r_1 and q_0 are time-independent; (2.2.2)

The proof of Theorem 2.2.1 is given in [8, Section 7.2].

2.3. Continuous observability. Global uniqueness. Dirichlet B.C.

Purely Dirichlet problem. Here we consider the following problem:

$$\left\{ \begin{array}{ll} iw_t + \Delta w = F(w) + f & \text{in } (0, T] \times \Omega \equiv Q; \\ w(0, \cdot) = w_0 & \text{in } \Omega; \end{array} \right. \quad (2.3.1a)$$

$$\left\{ \begin{array}{ll} w|_{\Sigma} = 0 & \text{in } (0, T] \times \Gamma = \Sigma, \end{array} \right. \quad (2.3.1c)$$

In the case of Dirichlet B.C. the extension Theorem 2.2.1 is *not* needed. Indeed the extension of the final sought-after continuous observability inequality (2.3.2) below can be readily accomplished from $H^{2,2}(Q)$ -solutions to $H^{1,1}(Q)$ -solutions just by virtue of the regularity theorem ('reverse inequality') obtained in [3] and reported also in [4]: the Neumann trace $\frac{\partial w}{\partial \nu}$ is dominated by the $H^1(\Omega)$ -norm of the I.C. w_0 . Thus the additional assumption (2.2.2) of Theorem 2.2.1 may be dispensed with. We obtain

Theorem 2.3.1. *Let w be a solution of problem (2.3.1) with I.C. $w_0 \in H_0^1(\Omega)$, and with $f \in L_2(0, T; H^1(\Omega))$, under the standing assumption (1.1.3) on $F(w)$, as well as (A.3) = (2.1.4). Assume, further, hypotheses (A.1)(ii) = (1.1.6), (A.2) = (1.1.7), for $d(x)$, see Remark 1.1.2. Define now Γ_0 as in (1.1.5_D), i.e., by: $h \cdot \nu \leq 0$, $h = \nabla d$, on Γ_0 . Let $\Gamma_1 = \Gamma \setminus \Gamma_0$. Let $T > 0$ be arbitrary. Then:*

(a) *there exists a constant $c_T > 0$, such that the following continuous observability inequality holds true:*

$$c_T E(0) \leq \int_0^T \int_{\Gamma_1} \left| \frac{\partial w}{\partial \nu} \right|^2 d\Gamma_1 dt + \|f\|_{L_2(0, T; H^1(\Omega))}^2. \quad (2.3.2)$$

The constant c_T is explicit, see [8].

(b) Let now $f = 0$ in (2.3.1). Then, the following global uniqueness result holds true: with $T > 0$, let w be a $H^{1,1}(Q)$ -solution of Eqn. (2.3.1a) and over-determined B.C.'s:

$$w|_{\Sigma} \equiv 0 \text{ and } \left. \frac{\partial w}{\partial \nu} \right|_{\Sigma_1} = 0, \text{ where } h \cdot \nu \leq 0, \text{ on } \Gamma_0 = \Gamma \setminus \Gamma_1, \quad (2.3.3)$$

$\Sigma_1 = (0, T] \times \Gamma_1$. Then, in fact, $w \equiv 0$ in Q , indeed in $\mathbb{R}_t \times \Omega$.

The proof of Theorem 2.3.1 is given in [8, Section 8].

2.4. Continuous observability.

Global uniqueness. Neumann B.C.

Purely Neumann problem. Here we consider the following problem:

$$\begin{cases} iw_t + \Delta w = F(w) + f & \text{in } (0, T] \times \Omega \equiv Q; \\ w(0, \cdot) = w_0 & \text{in } \Omega; \\ \left. \frac{\partial w}{\partial \nu} \right|_{\Sigma} = 0 & \text{in } (0, T] \times \Gamma \equiv \Sigma, \end{cases} \quad (2.4.1a) \quad (2.4.1b) \quad (2.4.1c)$$

As explained in more detail in [8, Section 7], it is for the Neumann B.C. case that the extension Theorem 2.2.1 is called for. Accordingly, we inherit its additional assumption (2.2.2).

Theorem 2.4.1. Let w be the solution of problem (2.4.1) with I.C. $w_0 \in H^1(\Omega)$, and with $f \in L_2(0, T; H^1(\Omega))$, under the standing assumption (1.1.3) on $F(w)$. Assume, further, hypotheses (A.1), (A.2), (A.3) [that is, (1.1.5_N), (1.1.6), (1.1.7), (2.1.4)], as well as the additional hypotheses as (A.4) = (2.2.2). Let $\Gamma_1 = \Gamma \setminus \Gamma_0$, where $h \cdot \nu = 0$ on Γ_0 as defined by (1.1.5_N), and let $T > 0$ be arbitrary. Then:

(a) there exists a constant $c_T > 0$, such that the following continuous observability inequality holds true:

$$c_T E(0) \leq \int_0^T \int_{\Gamma_1} [|w|^2 + |w_t|^2] d\Gamma_1 dt + \|f\|_{L_2(0, T; H^1(\Omega))}^2. \quad (2.4.2)$$

(b) Let now $f = 0$ in (2.4.1). Then, the following global uniqueness result holds true: with $T > 0$, let w be a $H^{1,1}(Q)$ -solution of Eqn. (2.4.1a) and over-determined B.C.'s:

$$\left. \frac{\partial w}{\partial \nu} \right|_{\Sigma} \equiv 0 \text{ and } w|_{\Sigma_1} \equiv 0, \text{ where } h \cdot \nu = 0 \text{ on } \Gamma_0 = \Gamma \setminus \Gamma_1, \quad (2.4.3)$$

$\Sigma_1 = (0, T] \times \Gamma_1$. Then, in fact, $w \equiv 0$ in Q , indeed in $\mathbb{R}_t \times \Omega$.

Indeed, [8] proves first the global uniqueness statement of part (b) of Theorem 2.4.1 in Section 8, as a direct consequence of the Carleman

estimate without lower-order terms of Theorem 2.1.2, first for smooth solutions and next extended to $H^{1,1}(Q)$ -solutions as in Theorem 2.2.1. Next, part (b) is used to establish part (a), in [8, Theorem 9.2 of Section 9] by virtue also of the trace Lemma 9.1 in [8].

Remark 2.4.1. Theorem 2.1.1 also yields a global uniqueness result in the purely Neumann case for solutions in the class (2.1.1).

2.5. Lower-level energy estimates

In this section, we formulate an estimate at a lower level of energy, that is, at the $L_2(\Omega)$ -level for w . To this end, we introduce anisotropic Sobolev spaces:

$$H_a^s(\Sigma) \equiv H^{s,s/2}(\Sigma) \equiv L_2(0, T; H^s(\Gamma)) \cap H^{s/2}(0, T; L_2(\Gamma)),$$

the latter Sobolev spaces being classical time-space spaces defined in [10, vol. II]. Then, $H_a^{-1}(\Sigma)$ is the dual space to $H_a^1(\Sigma)$, with respect to $L_2(\Sigma)$ as a pivot space. We then have in the prior notation:

Theorem 2.5.1. *Let w be a sufficiently smooth solution of the Schrödinger equation (1.1.1) with $f \in L_2(Q)$. Then, for any $T > 0$, the following inequality holds true: there exists a constant C_T such that*

$$\begin{aligned} & \int_0^T \left[\|w\|_{L_2(\Omega)}^2 + \|w_t\|_{H^{-2}(\Omega)}^2 \right] dt + \|w(0)\|_{L_2(\Omega)}^2 + \|w_t(0)\|_{H^{-2}(\Omega)}^2 \\ & \leq C_T \left\{ \|w\|_{L_2(\Sigma_1)}^2 + \left\| \frac{\partial w}{\partial \nu} \right\|_{H_a^{-1}(\Sigma)}^2 + \|w\|_{H^{-1}(Q)}^2 + \|f\|_{L_2(Q)}^2 \right\}. \end{aligned}$$

References

- [1] M. A. Horn and W. Littman, Boundary control of a Schrödinger equation with non-constant principal part, *Contr. Part. Diff. Eqns. Appl.*, E. Casas (ed.), Lecture Notes in Pure and Appl. Math. 174 (1996), 101–106.
- [2] M. Kazemi and M. V. Klibanov, Stability estimates for ill-posed Cauchy problems involving hyperbolic equations and inequalities, *Applicable Analysis* 50, (1993), 93–102.
- [3] I. Lasiecka and R. Triggiani, Optimal regularity, exact controllability and uniform stabilization of the Schrödinger equation, *Diff. Int. Eqns.* 5 (1991), 521–535.
- [4] I. Lasiecka and R. Triggiani, Carleman estimates and exact boundary controllability for a system of coupled, non-conservative second order hyperbolic equations, in Partial Differential Equations Methods in Control and Shape Analysis, *Lecture Notes in Pure & Applied Mathematics*, Marcel Dekker, G. Da Prato & J. P. Zolezio, editors, vol. 188, 215–243.

- [5] I. Lasiecka, R. Triggiani, and P. F. Yao, Inverse/observability estimates for second order hyperbolic equations with variable coefficients, *J. Math. Anal. & Appl.* 235 (1999), 13–57.
- [6] I. Lasiecka, R. Triggiani, and P. F. Yao, Carleman estimates for a plate equation on a Riemann manifold with energy level terms, Kluwer, ISAACS, to appear.
- [7] I. Lasiecka, R. Triggiani, and X. Zhang, Nonconservative wave equations with purely Neumann B.C.: Global uniqueness and observability in one shot, Amer. Math. Soc., *Cont. Math.*, 268 (2000), 227–326.
- [8] I. Lasiecka, R. Triggiani, and X. Zhang. Nonconservative Schrödinger equations with unobserved Neumann B.C. Global uniqueness and observability in one shot, preprint May 2002.
- [9] M. M. Lavrentev, V. G. Romanov, and S. P. Shishataskii, *Ill-Posed Prob. Math. Physics & Anal.*, Amer. Math. Soc., Vol. 64 (1986).
- [10] J. Lebeau, Contrôle de l'équation de Schrödinger, *J. Math. Pures & Appl.* 71 (1992), 267–291.
- [11] E. Machtyngier, Contrôlabilité exacte et stabilisation frontière de l'équation de Schrödinger, *C. R. Acad. Sc. Paris*, 310(I) (1990), 801–806.
- [12] D. Tataru, A-priori estimates of Carleman type in domains with boundary, *J. Math. Pure Appl.*, 73 (1994), 355–387.
- [13] D. Tataru, Boundary controllability of conservative PDEs, *Appl. Math. Optim.* 31 (1995), 257–295.
- [14] D. Tataru, Unique continuation for solutions to PDE's, between Hörmander's theorem and Holmgren's theorem, *Comm. Part. Diff. Eqns.* 20 (5 & 6) (1995), 855–884.
- [15] D. Tataru, Carleman estimates and unique continuation for solutions to boundary value problems, *J. Math. Pure Appl.* 75 (1996), 367–408.
- [16] R. Triggiani, Exact boundary controllability on $L_2(\Omega) \times H^{-1}(\Omega)$ of the wave equation with Dirichlet boundary control acting on a portion of the boundary, *Appl. Math. Optim.* 18 (1988), 241–277.
- [17] R. Triggiani, Carleman estimate and exact boundary controllability for a system of coupled non-conservative Schrödinger equations, special issue, *Rendi. dell' Istituto dei Matematica dell' Univ. di Trieste*, XXVIII (1996), 453–504, supplement dedicated to the memory of Pierre Grisvard.
- [18] R. Triggiani and P. F. Yao, Inverse/observability estimates for Schrödinger equations with variable coefficients, *Contr. & Cyber.* 28 (1999), 627–664, special issue on Control of Partial Differential Eqns.
- [19] R. Triggiani and P. F. Yao, Carleman estimates with no lower-order terms for general Riemann wave equations. Global uniqueness and observability in one shot, *Appl. Math. & Optimiz.*, to appear in December 2002.

OPTIMAL FLOW IN DYNAMIC NETWORKS WITH NONLINEAR COST FUNCTIONS ON EDGES

Dmitrii Lozovanu*

Institute of Mathematics and Computer Science

Moldovan Academy of Sciences

5 Academiei Str., Chisinau 2028

Republic of Moldova

lozovanu@math.md

Dan Stratila

Faculty of Mathematics and Computer Science

Moldova State University

60 Mateevici Str., Chisinau 2009

Republic of Moldova

dstrat@acm.org

Abstract We study the minimum-cost flow problem on dynamic networks with nonlinear cost functions that depend on time and edge flow. A general procedure for solving the problem using the time-expanded network is described. The main properties of dynamic flows on networks with concave cost functions are studied. We propose an algorithm for finding the optimal flow in networks with exactly one source; its running time is polynomial for a fixed number of sinks. Some details concerning the correctness of the algorithm and its computational complexity are discussed.

Keywords: dynamic networks, dynamic flows, flows over time, minimum-cost flows, concave cost functions.

*Supported by CRDF BGP MM2-3018.

1. Introduction and problem statement

In this paper we study the dynamic variant of the nonlinear minimum-cost flow problem on networks. This problem is a generalization of the classical static minimum-cost flow problem and is based on the dynamic network model from [1, 2]. We consider the problem on dynamic networks with nonlinear cost functions that depend on time and edge flow. We describe a procedure for solving the problem which is based on reducing the dynamic problem to the classical minimum-cost flow problem on a time-expanded network. This procedure is based on the approach from [1, 2] and represents a modification of the method in [3].

We show that in the case of uncapacitated networks with concave cost functions and infinite time horizon, as well as for a wide range of finite time bounds, there exists an optimal flow such that the edges used by the flow generate a forest. Using this result we propose a procedure that enables us to solve the dynamic problem on networks with concave cost functions on a static network of the same size. An algorithm for solving the problem on networks with one source is proposed; its running time is polynomial for a fixed number of sinks. Some details concerning the computational complexity of the algorithm and its correctness are discussed.

1.1. Flows in static networks

A *static network* $\mathcal{N} = (V, E, u, \varphi, d)$ consists of directed graph (V, E) , capacity function $u : E \rightarrow R_+$, demand and supply function $d : V \rightarrow R$ and cost function $\varphi : E \times R_+ \rightarrow R_+$. Nodes with $d_v < 0$ are called *sources*, nodes with $d_v > 0$ are called *sinks*, and nodes with $d_v = 0$ are called *intermediate*. We will denote by V_+, V_- , and V_* the sets of sources, sinks, and intermediate nodes respectively.

A *static flow* in \mathcal{N} [2] is a function $x : E \rightarrow \mathbb{R}_+$ that satisfies for all $v \in V$ the conservation constraints:

$$\sum_{e \in E_-(v)} x_e - \sum_{e \in E_+(v)} x_e = d_v. \quad (1)$$

Static flow x is called *feasible* if it satisfies capacity constraints $x_e \leq u_e$ for all $e \in E$. We note that in order for a flow to exist it is necessary for demand to equal supply: $\sum d_v = 0$.

The cost $\varphi(x)$ of static flow x is defined as:

$$\varphi(x) = \sum_{e \in E} \varphi_e(x_e).$$

The minimum-cost static flow problem is to find a feasible flow that minimizes the objective function $\varphi(x)$.

The graph $\mathcal{G}_x = (V_x, E_x)$ that consists of edge set $E_x = \{e | x_e > 0\}$ and node set $V_x = \{v | \exists w \text{ such that } (v, w) \in E_x \text{ or } (w, v) \in E_x\}$ is called the *base graph* of flow x in network \mathcal{N} .

1.2. Flows in dynamic networks

In the static model defined above flow travels across edges instantaneously. However, in many practical problems flow travel across edges may take non-zero time, and the cost of flow transit may change with time. If the time factor plays a significant role, dynamic networks and dynamic flows [1, 2] are often a better model.

A *dynamic network* $\mathcal{N} = (V, E, u, \tau, \varphi, d)$ consists of directed graph (V, E) , capacity function $u: E \rightarrow \mathbb{R}_+$, transit time function $\tau_e: E \rightarrow \mathbb{R}_+$, demand function $d: V \rightarrow \mathbb{R}$, and cost function $\varphi: E \times \mathbb{R}_+ \times \mathbb{T} \rightarrow \mathbb{R}_+$.

The meaning of τ_e is that flow entering edge $e = (v, w)$ at time t from node v will arrive at node w at time $t + \tau_e$. We consider the discrete time model, in which all times are integral and bounded by a *time horizon* T . The time horizon (finite or infinite) is the time until which flow can travel in the network and defines the *makespan* $\mathbb{T} = \{0, 1, \dots, T\}$ of time moments we consider.

To model transit costs, which may change over time, we define the cost function $\varphi_e(x_e(t), t)$, with the meaning that flow of value ξ entering edge e at time t will incur a transit cost of $\varphi_e(\xi, t)$. As with the static network, nodes are categorized into sources, sinks, and intermediate. Without losing generality, we will assume that no edges enter sources or exit sinks, and that $\varphi_e(0, t) = 0$ for all $e \in E$ and all $t \in \mathbb{T}$.

A *dynamic flow* is a function $x: E \times \mathbb{T} \rightarrow \mathbb{R}_+$ that satisfies the following constraints:

$$\sum_{e \in E_-(v)} \sum_{t=\tau_e}^{\theta} x_e(t - \tau_e) - \sum_{e \in E_+(v)} \sum_{t=0}^{\theta} x_e(t) \geq 0, \forall v \in V_*, \forall \theta \in \mathbb{T}; \quad (2)$$

$$\sum_{e \in E_-(v)} \sum_{t=\tau_e}^T x_e(t - \tau_e) - \sum_{e \in E_+(v)} \sum_{t=0}^T x_e(t) = d_v, \forall v \in V; \quad (3)$$

$$x_e(t) = 0, \forall e \in E, t = \overline{T - \tau_e + 1, T}. \quad (4)$$

The function defines the value $x_e(t)$ of flow entering edge e at time t . In order to obey the time horizon, flow must not enter an edge e at time t if it will have to leave the edge after time T , and this is ensured by constraint (4). As flow travels through the network, we allow unlimited

flow storage at the nodes, but prohibit any deficit by constraint (2). Finally, all demands must be met, flow must not remain in the network after time T , and each source must not exceed its supply. Thus is ensured by constraint (3).

Dynamic flow x is called feasible if it satisfies the capacity constraints $x_e(t) \leq u_e$ for all $e \in E$ and $t \in \mathbb{T}$. We note that in order for a flow to exist supply must equal demand: $\sum d_v = 0$.

The cost $\varphi(x)$ of dynamic flow x is defined as follows:

$$\varphi(x) = \sum_{e \in E} \sum_{t \in \mathbb{T}} \varphi_e(x_e(t), t).$$

Similarly to the static model, the dynamic minimum-cost flow problem is to find a feasible flow that minimizes the objective function $\varphi(x)$.

The graph $\mathcal{G}_x = (V_x, E_x)$ consisting of edge set $E_x = \{e \mid \sum_{t \in \mathbb{T}} x_e(t) > 0\}$ and node set $V_x = \{v \mid \exists w \text{ such that } (v, w) \in E_x \text{ or } (w, v) \in E_x\}$ is called the based graph of dynamic flow x in \mathcal{N} .

2. The time-expanded network

One of the most straightforward methods to tackle problems on a dynamic network $\mathcal{N} = (V, E, u, \tau, \varphi, d)$ is to reduce them to similar problems on a static *time-expanded* network $\mathcal{N}^T = (V^T, E^T, u^T, \varphi^T, d^T)$ defined as follows [2]:

- 1 $V^T := \{v(t) \mid v \in V, t \in \mathbb{T}\};$
- 2 $V_+^T := \{v(0) \mid v \in V_+\}, \text{ and } V_-^T := \{v(T) \mid v \in V_-\};$
- 3 $E^T := \{(v(t), w(t + \tau_e)) \mid e = (v, w) \in E, 0 \leq t \leq T - \tau_e\} \cup \{v(t), v(t + 1) \mid v \in V, 0 \leq t < T\};$
- 4 $u_{e(t)}^T := u_e, \text{ and } \varphi_{e(t)}^T(x_{e(t)}) := \varphi_e(x_e(t), t);$
- 5 $d_{v(t)}^T := d_v \text{ for } v(t) \in V_+^T \cup V_-^T, \text{ and } d_{v(t)}^T := 0 \text{ otherwise.}$

If we define a flow correspondence to be $x_{e(t)} := x_e(t)$ the minimum-cost flow problem on dynamic networks can be solved [2, 3, 4] by solving the static minimum-cost flow problem on the time-expanded network.

Unfortunately, for a dynamic network with n nodes, and m edges, we obtain a time-expanded network with $n(T + 1)$ nodes, and $O(mT)$ edges. In the case of $T = \infty$, we obtain an infinite time-expanded network \mathcal{N}^∞ . Since the size of the time-expanded network depends linearly on T , direct use of this method does not lead to polynomial algorithms, and for large T yields time-expanded networks that are not practical to work with.

However, Fleischer and Skutella in [4] present polynomial approximation algorithms that employ “condensed” time-expanded networks which rely on a rougher discretization of time. In [3] it is shown that for an acyclic network with unit-time transit times and zero flow storage at nodes it is sufficient to consider a time-expanded network of size at most $O(n^2)$ nodes and $O(mn)$ edges for any T , finite or infinite.

In the following section we present a different approach for uncapacitated dynamic networks with cost functions that are concave with regard to flow value, and do not change over time. Relying on concavity, we reduce the problem to the minimum-cost flow problem on a static network of equal size, not the time-expanded network.

3. Networks with concave cost functions

Most previous work involving flow costs in dynamic networks considers linear [4, 5] or convex [6, 7] cost functions with regard to flow value. This implies that the cost of transporting one unit of flow is the same regardless of how many units are transported at a time, or that the cost per unit is rising with the total number of units. However, in many practical cases transports of larger quantities enable discounts on the price per unit. This behaviour is best modeled by cost functions concave with regard to flow value.

3.1. Flow properties

Given a dynamic network $\mathcal{N} = (V, E, u, \tau, \varphi, d)$, the corresponding static network \mathcal{N}^0 of \mathcal{N} is obtained by discarding all time-related information: $\mathcal{N}^0 = (V, E, u, \varphi^0, d)$, where $\varphi_e^0(\xi) = \varphi_e(\xi, 0)$.

Lemma 1. *Let \mathcal{N} be an uncapacitated dynamic network with cost functions concave with regard to flow and constant in time. If x is a flow in \mathcal{N} , then $y_e = \sum_{t \in \mathbb{T}} x_e(t)$ is a flow in the corresponding static network \mathcal{N}^0 and $\varphi^0(y) \leq \varphi(x)$.*

Proof. Note that if $\phi : R_+ \rightarrow R_+$ is a concave function, then $\phi(\alpha + \beta) \leq \phi(\alpha) + \phi(\beta)$ for all $\alpha, \beta \in R_+$. Since φ and φ^0 are concave with regard to flow value, we obtain $\varphi^0(y) = \sum_{e \in E} \varphi_e^0(y_e) = \sum_{e \in E} \varphi_e^0(\sum_{t \in \mathbb{T}} x_e(t)) \leq \sum_{e \in E} \sum_{t \in \mathbb{T}} \varphi_e^0(x_e(t)) = \sum_{e \in E} \sum_{t \in \mathbb{T}} \varphi_e(x_e(t), t) = \varphi(x)$. Moreover, $y_e = \sum_{t=0}^T x_e(t) = \sum_{t=\tau_e}^T x_e(t - \tau_e)$, since flow x obeys constraint (4). By substituting y_e in dynamic conservation constraint (3), we obtain the static conservation constraint (1). Therefore y is a flow in \mathcal{N}^0 . ■

This lemma relies on the fact that in uncapacitated networks each dynamic flow induces a static flow in the corresponding static network.

While the converse is not true in the general case, it holds for flows with forest base graphs and sufficiently large time horizons.

Lemma 2. *Let \mathcal{N} be an infinite-horizon dynamic network with cost functions constant in time. If y is a static flow in \mathcal{N}^0 such that its base graph \mathcal{G}_y is a forest, then there exists a dynamic flow x in \mathcal{N} such that $\varphi(x) = \varphi^0(y)$.*

Proof. Let $x_{(v,w)}(t) = y_{(v,w)}$ if $t = t_v$, and $x_{(v,w)}(t) = 0$ otherwise, where:

$$t_v = \begin{cases} 0, & \text{if } v \in V_+, \\ \max\{t_w + \tau_{(w,v)} | (w, v) \in E_y\}, & \text{otherwise.} \end{cases} \quad (5)$$

Since $\mathcal{G}_y = (V_y, E_y)$ is a forest, the constants t_v are well-defined and finite. To prove that x is a flow in \mathcal{N} , we have to show that it satisfies constraints (4), (2), and (3).

Because $T = +\infty$ it follows that $T \geq t_v, \forall v \in V_y$. Therefore, for any $e = (v, w) \in E_y$, we obtain $T \geq t_w \geq t_v + \tau_e$, hence $T - \tau_e \geq t_v$. Since $t \neq t_v$ implies $x_e(t) = 0$, it follows that $x_e(t) = 0$ for all $t_v > T - \tau_e \geq t_v$, hence constraint (4) is obeyed.

Definition (5) means flow starts leaving $\forall v \in V_* \cap V_y$ only after all inbound flow has arrived. Thus for $\theta < t_v$ we have $\sum_{e \in E_+(v)} \sum_{t=0}^\theta x_e(t) = 0$, hence constraint (2) holds for $\theta < t_v$. For $\theta \geq t_v$ flow summed over time on any edge is the same as the flow on that edge in the static network: $\sum_{e \in E_-(v)} \sum_{t=\tau_e}^\theta x_e(t - \tau_e) = \sum_{e \in E_-(v)} y_e = \sum_{e \in E_+(v)} y_e = \sum_{e \in E_+(v)} \sum_{t=0}^\theta x_e(t)$. Therefore, constraint (2) holds for $\theta \geq t_v$. We have established that constraint (2) is obeyed.

By taking $\theta = T \geq t_v$ in the previous argument, we obtain that constraint (3) holds for all $v \in V_* \cap V_y$. For all sources $v \in V_+$ incoming flow is zero: $\sum_{e \in E_-(v)} \sum_{t \in \mathbb{T}} x_e(t) = 0$, since no edges enter a source. On the other hand, outgoing flow equals supply: $\sum_{e \in E_+(v)} \sum_{t \in \mathbb{T}} x_e(t) = \sum_{e \in E_+(v)} y_e = -d_v$. Therefore, constraint (3) holds for all sources. The proof for sinks is similar, taking into account no edges exit sinks. Therefore, constraint (3) is obeyed.

Having proven that x is a flow, it is easy to see that it is feasible, since $0 \leq x_e(t) \leq y_e \leq u_e$. Finally, $\varphi(x) = \sum_{e \in E} \sum_{t \in \mathbb{T}} \varphi_e(x_e(t), t) = \sum_{e \in E_y} \varphi_e(x_e(t_v), t_v) = \sum_{e \in E} \varphi_e^0(y_e) = \varphi^0(y)$. ■

In the above proof we employ the fact that $T = +\infty$ only to maintain that $t_v \leq T, \forall v \in V_y$. However, if we denote by $|L| = \sum_{e \in L} \tau_e$ the time-length of a path in \mathcal{N} , then we immediately obtain that $t_v \leq \max_{L \in \mathcal{N}} \{|L|\}$. Hence $\max_{L \in \mathcal{N}} \{|L|\}$ is an upper bound for the makespan

of flow x as constructed in the above lemma, and we can broaden the class of networks we examine.

Lemma 3. *Let \mathcal{N} be a dynamic network with cost functions constant in time such that $T \geq \max_{L \in \mathcal{N}}\{|L|\}$. If y is a static flow in \mathcal{N}^0 such that its base graph \mathcal{G}_y is a forest, then there exists a dynamic flow x in \mathcal{N} such that $\varphi(x) = \varphi^0(y)$.*

To make the connection between Lemma 1, Lemma 3, and minimum-cost flows in dynamic networks, we will employ the following property of minimum-cost flows in static networks with concave cost functions [8, 9].

Lemma 4. *Let \mathcal{N}^0 be an uncapacitated static network with concave non-decreasing cost functions. If there exists a flow in \mathcal{N}^0 , then there exists a minimum-cost flow y in \mathcal{N}^0 such that its base graph \mathcal{G}_y is a forest.*

We are now able to prove the main result of this sub-section. Denote by y^T the dynamic flow in \mathcal{N} obtained from a forest-like flow y in \mathcal{N}^0 such that $y_{(v,w)}^T(t) = y_{(v,w)}$ if $t = t_v$, and $y_{(v,w)}^T(t) = 0$ otherwise, where t_v are defined as in (5).

Theorem 1. *Let \mathcal{N} be an uncapacitated dynamic network with cost functions concave with regard to flow and constant in time such that $T \geq \max_{L \in \mathcal{N}}\{|L|\}$. If there exists a flow in \mathcal{N} , then there exists a minimum-cost forest-like flow z in \mathcal{N}^0 , and the flow z^T is a minimum-cost flow in \mathcal{N} .*

Proof. Since there exists a flow in \mathcal{N} , a flow can be constructed in \mathcal{N}^0 according to Lemma 1. Hence, according to Lemma 4 there exists a minimum-cost forest-like flow in \mathcal{N}^0 ; denote this flow by z . Flow z^T is a minimum-cost flow in \mathcal{N} . Indeed, for any flow x in \mathcal{N} we have $\varphi(z^T) = \varphi^0(z) \leq \varphi^0(y) \leq \varphi(x)$, where y is a static flow in \mathcal{N}^0 such that $y_e = \sum_{t \in \mathbb{T}} x_e(t)$. Equality $\varphi(z^T) = \varphi^0(z)$ follows from Lemma 3, inequality $\varphi^0(z) \leq \varphi^0(y)$ from the fact that z is a minimum-cost flow in \mathcal{N}^0 , and inequality $\varphi^0(y) \leq \varphi(x)$ from Lemma 1. ■

Therefore, a dynamic minimum-cost flow can be computed using the following procedure: (i) find a forest-like minimum-cost flow z in the corresponding static network; (ii) compute the constants t_v and construct the dynamic minimum-cost flow z^T .

3.2. An algorithmic approach

We will present a combinatorial algorithm for the minimum-cost flow problem on dynamic networks that meet the conditions of Theorem 1, and have exactly 1 source, based on the approach in [9]. We begin by

examining networks with 1, 2, or 3 sinks, and then present an algorithm for the general case with k sinks. Without loss of generality we will label the source node 0, and the k sinks $0, 1, \dots, k$.

We seek a minimum-cost flow with a base graph that is a forest. Since there is exactly 1 source, and $\sum d_v = 0$, it follows that the base graph is connected, and thus it is a tree. Furthermore, since there are no edges entering sources or exiting sinks, the base graph rooted at the source will have the sinks as leaves; the problem becomes related to network design problems and Steiner trees [10].

One sink. We seek a rooted tree \mathcal{G}_y with one leaf, in other words a simple path from the source 0 to sink 1. Obviously the flow on any edge of the path is equal to the total demand d_1 , and we need to minimize $\sum_{e \in \mathcal{G}_y} \varphi_e^0(d_1)$. This can be achieved by finding a shortest path from 0 to 1 with regard to edge lengths $\varphi_e^0(d_1)$. The coefficients t_v for reconstructing the dynamic flow can be computed trivially by parsing the path from source to sink.

Two sinks. We seek a tree with fixed root 0 (the source) and two fixed leaves 1 and 2 (the sinks). Any such tree would have the structure of one of the trees in Figure 1, where by arrows we have denoted either simple paths or edges. Moreover tree structure (a) contains the other tree structures as degenerate cases; hence we can assume all trees have this structure. Denote by d_W demand $\sum_{v \in W} d_v$; then flow on the paths from 0 to w , w to 1, and w to 2 equals d_{12} , d_1 , and d_2 respectively.

Consequently, we need to minimize for all $w \in V$ flow cost $\varphi^0(y) = \sum_{e \in L_{0w}} \varphi_e^0(d_{12}) + \sum_{e \in L_{w1}} \varphi_e^0(d_1) + \sum_{e \in L_{w2}} \varphi_e^0(d_2)$. This can be achieved with 3 one-to-all shortest path computations. Denote by c_{vw}^W the minimum distance from v to w with regard to edge lengths $\varphi^0(d_W)$; we compute for all $w \in V$ c_{0w}^{12} from 0 to w , c_{w1}^1 backwards from 1 to w , and c_{w2}^2 backwards from 2 to w . Then it is straightforward to find $\min_{w \in V} \{c_{0w}^{12} + c_{w1}^1 + c_{w2}^2\}$ and then to compute t_v by parsing the resulting tree. Due to concavity path overlapping does not lead to invalid results; we will address this in the proofs for the general case.

Three sinks. We seek a tree with fixed root 0 and fixed sink leaves 1, 2, and 3. There are more structures such trees could take, but all are degenerate or proper cases of the three structures depicted in Figure 2. To find a minimum-cost tree-like flow, we compute a minimum-cost flow tree for each structure, and then select the optimal one. Consider computing a minimum-cost flow tree that fits structure (c) from Figure 2. Its cost can be expressed as the sum of costs along tree branches, as in the case

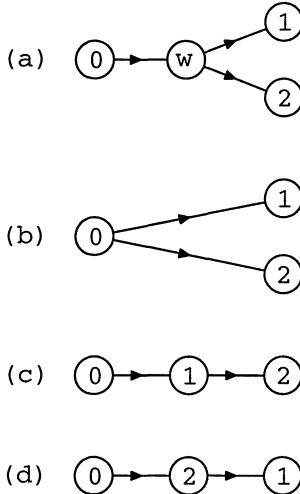


Figure 1. All structures for trees with one source and 2 sinks. Structure (b) is a degenerate case obtained from (a) by setting $w = 0$; structure (c) is obtained from (a) by setting $w = 1$; structure (d) is obtained from (a) by setting $w = 2$.

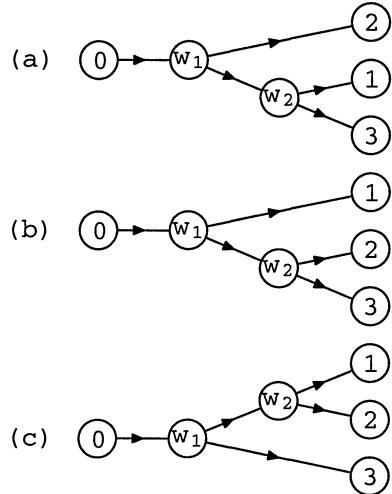


Figure 2. All non-degenerate structures for trees with one source and 3 sinks. All other structures can be obtained as degenerate cases by using various combinations of setting w_1 and w_2 to 0, 1, 2, 3, and each other.

with two sinks. Therefore, the minimization can be achieved with $O(n)$ one-to-all shortest path computations. We compute $c_{v1}^1, c_{v2}^2, c_{v3}^3, c_{0v}^{123}$ for all $v \in V$, as well as c_{vw}^{12} for all $v \in V$ and $w \in V$. Then we find $\min_{v,w \in V} \{c_{0v}^{123} + c_{vw}^{12} + c_{v3}^3 + c_{w1}^1 + c_{w2}^2\}$ and finally t_v . Minimum-cost trees for the other two structures are computed similarly.

The general case. We use dynamic programming to generalize, and present algorithm **MINFLOW** for computing the minimum cost of a dynamic flow in network \mathcal{N} that satisfies the conditions of Theorem 1. We will consider given: (i) the dynamic network \mathcal{N} with n nodes, m edges, 1 source, and k sinks; (ii) the set of all tree structures with k leaves \mathbb{A} .

Each structure $\mathcal{A} = (V_{\mathcal{A}}, E_{\mathcal{A}}) \in \mathbb{A}$ that fits a tree G_x is represented itself as a tree that contains the root and all sinks in G_x . Moreover, \mathcal{A} contains a new node for each node in G_x that has more than one outbound edge; we will call these *branch* nodes. Tree \mathcal{A} also contains a new edge for each simple path in G_x consisting of nodes with exactly one outbound and exactly one inbound edge.

The algorithm then calls **MINTREE** for each structure $\mathcal{A} \in \mathbb{A}$, which returns the minimum-cost of a flow with structure \mathcal{A} . The key element of **MINTREE** is the matrix $c[v][w]$, where $v \in V_{\mathcal{A}}$ and $w \in V$. The algorithm computes all elements of this matrix so that $c[v][w]$ equals the cost of sending $d[v]$ units of flow from node w to the sinks that are in the subtree \mathcal{A}_v of node v so that to satisfy their demands. Here $d[v] = d_v$ for all sinks $v \in V_-$ and $d[v] = \sum_{\alpha \in \mathcal{A}_v \cap V_-} d_\alpha$ for all branch nodes. More loosely, $c[v][w]$ represents the cost of “assigning” branch node v to network node w .

For all sinks $v \in V_-$ the algorithm initializes $c[v][v]$ to 0; indeed the cost of getting the flow to v from v is zero. All other elements $c[v][w]$ with $v \in V_-$ and $v \neq w$ are assigned $+\infty$. Then we are able to efficiently compute $c[v][w]$ by walking up the tree \mathcal{A} , and in the end obtain the sought minimum cost in element $c[0][0]$.

MINFLOW(\mathcal{N}, \mathcal{A}):

```

 $c_{\min} := +\infty;$ 
for each  $\mathcal{A} \in \mathbb{A}$  do:
   $c := \text{MINTREE}(\mathcal{N}, \mathcal{A});$ 
  if  $c < c_{\min}$  then  $c_{\min} := c$ ;
end for;
return  $c_{\min}$ .

```

MINLEN($\mathcal{N}, v, l[]$):

```

return minimum distances  $s[]$  from  $v$  to
all nodes in the network  $\mathcal{N}$  with regard
to edge lengths  $l[]$ .

```

DESC(\mathcal{A}, v):

```

return the set  $E_{\mathcal{A}}^+(v)$  of all direct descendants
of node  $v$  in tree  $\mathcal{A}$ .

```

READY($\mathcal{A}, b[]$):

```

return the set  $\{v \in V_{\mathcal{A}} | b[v] = 0 \wedge b[w] =$ 
 $1, \forall w \in \text{DESC}(\mathcal{A}, v)\}$  of all unprocessed
nodes ready to be processed.

```

MINTREE(\mathcal{N}, \mathcal{A}):

```

for each  $v \in V_{\mathcal{A}}$  do:
   $b[v] := 0;$ 
   $c[v][] := +\infty;$ 
end for;
while  $\exists v \in \text{READY}(\mathcal{A}, b[])$  do:
   $b[v] := 1;$ 
  if  $v \notin V_-$  then
     $d[v] := \sum_{w \in \text{DESC}(\mathcal{A}, v)} d[w];$ 
  for each  $e \in E$  do
     $l[v][e] := \varphi_e(d[v], 0);$ 
  for each  $w \in V$  do:
    if  $v \notin V_-$  or  $v = w$  then
       $c[v, w] := 0;$ 
    for each  $\alpha \in \text{DESC}(\mathcal{A}, v)$  do:
       $s[] := \text{MINLEN}(\mathcal{N}, w, l[\alpha]);$ 
       $c[v, w] := c[v, w]$ 
        +  $\min_{\beta \in V} \{s[\beta] + c[\alpha, \beta]\};$ 
    end for;
  end for;
end while;
return  $c[0, 0]$ .

```

The algorithm can be modified to return not just the minimum cost, but also the flow for which it is obtained.

Lemma 5. *Let \mathcal{N} be an uncapacitated dynamic network with cost functions concave with regard to flow and constant in time such that $T \geq \max_{L \in \mathcal{N}} \{|L|\}$. If there exists a flow in \mathcal{N} , then the cost computed by sub-algorithm **MINTREE**(\mathcal{N}, \mathcal{A}) equals or exceeds the cost of at least one flow in \mathcal{N} .*

Proof. It results from the definition of **READY** and the use of vector b that at the point when the algorithm begins computing $c[v][]$ for a non-sink node $v \in E_A$, $c[w][]$ is already computed for all descendants $w \in E_A^+(v)$. It also holds that $d[v]$ is computed to equal the total demand of all sinks under v .

When computing $c[v][w]$, the algorithm sums for each descendant the cost of sending flow from that descendant plus the cost of getting the flow to that descendant. On the other hand, due to concavity, the sum of the costs of sending ξ_1 and ξ_2 units of flow across an edge e equals or exceeds the cost of sending $\xi_1 + \xi_2$ units together.

It follows by induction up the tree A that $c[v, w]$ equals or exceeds the cost of at least one way of sending $d[v]$ units of flow from node w to the sinks in the sub-tree with root v in A while obeying flow conservation constraints and meeting sink demands. Therefore $c[0][0]$ equals or exceeds the cost of at least one way of sending $-d_0$ units of flow from source 0 to all the sinks while meeting their demands and obeying flow conservation constraints. ■

Lemma 6. *Let \mathcal{N} be an uncapacitated dynamic network with cost functions concave with regard to flow and constant in time such that $T \geq \max_{L \in \mathcal{N}}\{|L|\}$. If there exists a minimum-cost tree-like flow x with structure A in \mathcal{N} , the cost computed by **MINTREE**(\mathcal{N}, A) equals the minimum cost of a flow in \mathcal{N} .*

Proof. If flow x has structure A then we can associate to each node $v \in A$ a node $v_x \in V$. We will refer to the cost of flow being transported on the edges of a sub-tree of the base graph \mathcal{G}_x as the cost of that subtree.

We note that **MINTREE** computes the shortest path from a node to its descendants, and uses the minimum total cost selecting each descendant. It follows by induction up the tree A that $c[v, v_x]$ is less than or equal to the cost of the subtree rooted at v_x with regard to flow x .

Therefore, $c[0][0] \leq \varphi(x)$. But according to Lemma 5, $c[0][0]$ exceeds or equals the cost of a flow in \mathcal{N} . Since y is a minimum-cost flow in \mathcal{N} we obtain that $c[0][0] = \varphi(x)$. ■

Theorem 2. *Let \mathcal{N} be an uncapacitated dynamic network with cost functions concave with regard to flow and constant in time such that $T \geq \max_{L \in \mathcal{N}}\{|L|\}$. If there is a flow in \mathcal{N} , then algorithm **MINFLOW** computes the minimum cost of a flow in the dynamic network \mathcal{N} .*

Proof. Since there is a flow in \mathcal{N} , there exists a minimum-cost flow x in \mathcal{N} , and hence there is a structure $A \in \mathbb{A}$ that fits it. Therefore, according to Lemma 6 **MINTREE**(\mathcal{N}, A) will return the minimum cost of a flow in \mathcal{N} . Moreover, since according to Lemma 5 any cost returned by **MINTREE** equals or exceeds the cost of a flow, all other calls

will return costs equal to or higher than the cost of the minimum flow. Since **MINFLOW** selects the minimum from all calls, it will return the minimum cost of a flow in \mathcal{N} . ■

3.3. Algorithm complexity

We first examine the complexity of sub-algorithm **MINTREE**(\mathcal{N}, A). With an appropriate choice of tree representation, the calls to **READY** and **DESC** do not impact the order of complexity given by the loop structure. We assume Dijkstra's algorithm is used for the **MINLEN** computations. From the loop structure of **MINTREE** it follows that its execution has the complexity $O(|V_A| + |V_A|(m+n(n \log n + m + n))) = O(kn(n \log n + m))$. Therefore algorithm **MINFLOW** has complexity $O(|\mathbb{A}|kn(n \log n + m))$. Obviously, the number of tree structures in \mathbb{A} is at least exponential of the number of leaves k . However, if we consider networks with k fixed (not part of the input), then the algorithm is polynomial, with complexity $O(mn + n^2 \log n)$.

References

- [1] L. R. Ford, Jr. and D. R. Fulkerson. Constructing maximal dynamic flows from static flows. *Operations Res.*, 6:419–433, 1958.
- [2] L. R. Ford, Jr. and D. R. Fulkerson. *Flows in networks*. Princeton University Press, Princeton, N.J., 1962.
- [3] D. Lozovanu and D. Stratila. The minimum-cost flow problem on dynamic networks and an algorithm for its solving. *Bul. Acad. Științe Repub. Mold. Mat.*, (3):38–56, 2001.
- [4] Lisa Fleischer and Martin Skutella. The quickest multicommodity flow problem. In *Integer Programming and Combinatorial Optimization*, pages 36–53. Springer, Berlin, 2002.
- [5] Bettina Klinz and Gerhard J. Woeginger. Minimum cost dynamic flows: the series-parallel case. In *Integer programming and combinatorial optimization (Copenhagen, 1995)*, pages 329–343. Springer, Berlin, 1995.
- [6] James B. Orlin. Minimum convex cost dynamic network flows. *Math. Oper. Res.*, 9(2):190–207, 1984.
- [7] Lisa Fleischer and James B. Orlin. Optimal rounding of instantaneous fractional flows over time. *SIAM J. Discrete Math.*, 13(2):145–153 (electronic), 2000.
- [8] D. D. Lozovanu. Properties of optimal solutions of a grid transport problem with concave cost function of the flows on the arcs. *Izv. Akad. Nauk SSSR Tekhn. Kibernet.*, (6):94–98, 1982.
- [9] D. D. Lozovanu. *Ekstremalno-kombinatornye zadachi i algoritmy ikh resheniya*. “Shtiintsa”, Kishinev, 1991.
- [10] Frank K. Hwang, Dana S. Richards, and Pawel Winter. *The Steiner tree problem*. North-Holland Publishing Co., Amsterdam, 1992.

METHOD OF EXTREMAL SHIFT IN PROBLEMS OF RECONSTRUCTION OF AN INPUT FOR PARABOLIC VARIATIONAL INEQUALITIES

Vyacheslav Maksimov*

*Institute of Mathematics and Mechanics,
Ural Branch, Russian Academy of Sciences,
S. Kovalevskoi str., 16, Ekaterinburg, 620219 Russia
maksimov@imm.uran.ru*

Abstract The problem of dynamical reconstruction of a control in a parabolic variational inequality [1–4] through inaccurate measurement of phase state is considered. This problem belongs to the class of dynamical inverse problems which consist in reconstruction of unknown input (a control) of dynamical systems from (inaccurate) measurement of outputs.

Keywords: reconstruction, parabolic inequality

Introduction

Let V and H be real Hilbert spaces, V be a dense subspace of H and $V \subset H = H^* \subset V^*$ algebraically and topologically. We denote by $|\cdot|_V$ and $|\cdot|_H$ the norms in V and H , respectively, by (\cdot, \cdot) a scalar product in H , and by $\langle \cdot, \cdot \rangle$ duality between V and its dual V^* . Let a dynamical system be described by the parabolic variational inequality

$$(x_t(t) - f(t), x(t) - z) + \langle Ax(t) - Bu(t), x(t) - z \rangle + \varphi(x(t)) - \varphi(z) \leq 0 \quad (1)$$

a. e. $t \in T = [t_0, \vartheta]$ $\forall z \in V$, $x(t_0) = x_0 \in D(\varphi)$.

Here, $A : V \rightarrow V^*$ is a linear continuous symmetrical operator satisfying (with a certain $c > 0$ and $\omega \in R$) the following condition

$$\langle Ay, y \rangle + \omega |y|_H^2 \geq c |y|_V^2; \quad (2)$$

*This work was supported in part by the Russian Foundation for Basic Research (grant # 01-01-00566).

$f(\cdot) \in W^{1,2}(T; H)$ is a given disturbance; $\varphi : V \rightarrow R^+ = \{r \in R : r \geq 0\} \cup \{+\infty\}$ is a weakly lower semicontinuous convex function, $D(\varphi) = \{x \in V : \varphi(x) < +\infty\}$, U is a uniformly convex Banach space (the space of controls), $W^{1,2}(T; H) = \{x(\cdot) \in L_2(T; H) : x_t(\cdot) \in L_2(T; H)\}$, the derivative $x_t(\cdot)$ is understood in a sense of distribution; B is a linear continuous operator from U to H .

Introduce a function $\varphi_\omega(y) : H \rightarrow R^+$,

$$\varphi_\omega(y) = \begin{cases} 1/2\langle Ay, y \rangle + \omega/2|y|_H^2 + \varphi(y), & \text{if } y \in D(\varphi), \\ +\infty & \text{otherwise.} \end{cases}$$

From Theorem 4.1, 1.13 [2] it follows that for any $u(\cdot) \in L_2(T, U)$, $x_0 \in D(\varphi)$ there exists a unique solution of the inequality (1) with the properties: $x(\cdot) = x(\cdot; t_0, x_0, u(\cdot)) \in W_*(T) = W^{1,2}(T; H) \cap L_2(T; V)$, $x(t) \in D(\varphi_\omega) \quad \forall t \in T, \quad t \rightarrow \varphi_\omega(x(t)) \in AC(T)$. Here $AC(T)$ is a space of absolutely continuous functions.

The problem may be formulated in the following way. Let a uniform net $\Delta = \{\tau_i\}_{i=0}^m$, $\tau_i = \tau_{i-1} + \delta$, $\tau_0 = t_0$, $\tau_m = \vartheta$ be fixed on a given time interval T . Let a motion of system (1) proceed on the interval T . Its trajectory (the solution of (1)) $x_r(\cdot) = x(\cdot; t_0, x_0, u_r(\cdot))$ depends on time-varying unknown control $u(\cdot) = u_r(\cdot) \in U_T = L_2(T; U)$. Here U_T is a set of admissible controls. The trajectory $x_r(\cdot)$ is unknown. At moments τ_i the phase state $x_r(\tau_i)$ is inaccurately measured. The results of measurements $\xi_i^h \in H$, $i \in [0 : m - 1]$ satisfy the inequality

$$|\xi_i^h - x_r(\tau_i)|_H \leq h. \quad (3)$$

Here, $h \in (0, 1)$ is the value of the level of informational noise. It is required to design an algorithm allowing to reconstruct (synchro with the process) some unknown input $u_*(\cdot)$ generating solution $x_r(\cdot)$ to inequality (1). This is meaningful statement of the problem.

1. The approach to solving the problem

One of the approaches to solving the problems of similar type was described, for example, in [5–12]. (See more detail on this approach in surveys [13–16]). Briefly remind about the essence of this approach. Let $U(x_r(\cdot))$ be the set of all inputs $u(\cdot) \in U_T$ which are compatible with $x_r(\cdot)$, Ξ_T be the set of all measurements, i. e. the set of all piece-wise functions on T with values in H , $\Xi(x_r(\cdot), h)$ be the set of all h -accurate measurements, i. e. functions $\xi^h(\cdot) \in \Xi_T$ satisfying (3). An auxiliary system M (a model) described by a parabolic variational inequality of the form

$$(w_t^h(t) - f(t), w^h(t) - z) + \langle Aw(t) - B_t(w^h(\cdot), \xi^h(\cdot)), v^h(t), w^h(t) - z \rangle + \varphi(w^h(t)) - \varphi(z) \leq 0 \text{ a. e. } t \in T \quad \forall z \in V, \quad w^h(t_0) = w_0^h \in D(\varphi) \quad (4)$$

is considered. Here $B_t(\cdot, \cdot) : U \rightarrow H$ is a family of operators, whose structure will be specified below. The motion of the model is denoted as $w^h(\cdot) = w^h(\cdot; t_0, w_0^h, v^h(\cdot)) \in W_*(T)$. Here $v^h(\cdot)$ is a control in the model. Initial state w_0^h of the model is chosen by using the value ξ_0^h of measurement at initial time moment t_0 in accordance with some rule \mathcal{W}_h fixed in advance:

$$w_0^h = w^h(t_0) = \mathcal{W}_h(\xi_0^h) \in H. \quad (5)$$

Model control rules are identified with pairs $S_h = (\Delta_h, \mathcal{U}_h)$, where $\Delta_h = \{\tau_{h,i}\}_{i=0}^{m_h}$ is a partition of the interval T into half-intervals $[\tau_{h,i}, \tau_{h,i+1})$, $\tau_{h,i+1} = \tau_{h,i} + \delta$, $\delta = \delta(h)$, $\tau_{h,0} = t_0$, $\tau_{h,m_h} = \vartheta$, \mathcal{U}_h is a function relating element

$$v_i^h = \mathcal{U}_h(\tau_i, \xi_i^h, w^h(\tau_i)) \in U \quad (6)$$

to every triple $(\tau_i, \xi_i^h, w^h(\tau_i))$, $i \in [0 : m_h - 1]$, where $\tau_i = \tau_{h,i}$, $w^h(\tau_i) = w^h(\tau_i; t_0, w_0^h, v^h(\cdot))$, $\xi_i^h = \xi^h(\tau_i)$, $\xi^h(\cdot) \in \Xi(x_r(\cdot), h)$. Thus, quadruple $(M, \mathcal{W}_h, \Delta_h, \mathcal{U}_h)$ for every $h \in (0, 1)$ determines some algorithm D_h on space of measurements $\xi(\cdot) \in \Xi(x_r(\cdot), h)$ ($D_h : \Xi_T \rightarrow U_T$) forming output $v^h(\cdot) = D_h \xi(\cdot)$ according to the feedback principle (4)–(6). We identify the algorithm D_h with quadruple $(M, \mathcal{W}_h, \Delta_h, \mathcal{U}_h)$. The result of work of the algorithm on the interval T is a piecewise constant control $v^h(\cdot)$ of the form

$$v^h(t) = v_i^h, \quad t \in [\tau_i, \tau_{i+1}). \quad (7)$$

Let the following condition be fulfilled.

Condition 1 The set $U_*(x_r(\cdot))$ of $L_2(T; U)$ -norm minimal inputs in $U(x_r(\cdot))$ is one-element, i. e. $U_*(x_r(\cdot)) = \{u_*(\cdot; x_r(\cdot))\}$.

A family D_h , $h \in (0, 1)$ of operators from Ξ_T to U_T is called *regularizing* if

$$\lim_{h \rightarrow 0} \sup \{|D_h \xi^h(\cdot) - u_*(\cdot; x_r(\cdot))|_{L_2(T; U)} : \xi^h(\cdot) \in \Xi(x_r(\cdot), h)\} = 0.$$

We consider the problems of construction of regularizing families of algorithms of modeling

$$D_h = (M, \mathcal{W}_h, \Delta_h, \mathcal{U}_h), \quad h \in (0, 1) \quad (8)$$

of the form (4)–(7). We call them *positional algorithms of modeling*.

For parabolic variational inequality in [9–11], some regularizing families of algorithms of modeling D_h of type (8) were indicated for the case when the set of admissible controls U_T has the form: $U_T = \{u(\cdot) \in L_2(T; U) : u(t) \in P \text{ for a.e. } t \in T\}$. Here $P \subset U$ is a convex, bounded, and closed set. The constructions of works [9–11] are based on combination of the theory of control with a model [17] and the known in

the theory of ill-posed problems method of smoothing functional (also often called Tikhonov's method). In the present work, we modify algorithms [5–8, 10–12] for the case when set U_T is unbounded. Note that works [9, 13] indicate some other regularizing algorithms D_h of reconstruction of unbounded controls in parabolic variational inequalities based on a dynamical modification of the discrepancy method. Questions on program control of parabolic variational inequalities have been discussed in monographs [2].

After model (4) and its initial state (5) are chosen the work of the algorithm D_h (for fixed h) corresponds to the following outline. First, before the moment t_0 , a partition $\Delta = \Delta_h = \{\tau_i\}_{i=0}^m$, ($\tau_i = \tau_{h,i}$) of the interval T is chosen and fixed. At the i -th step carried out during the time interval $[\tau_i, \tau_{i+1})$, the following sequence of actions takes place. The output $x_r(\tau_i)$ is inaccurately measured, i. e. the value $\xi_i^h \in H$ with the properties (3) is calculated. Then the model control is determined by (6), (7) and after that we form the new part of the model trajectory $w^h(t)$, $t \in (\tau_i, \tau_{i+1}]$ instead of $w_{t_0, \tau_i}^h(\cdot)$ (memory correction). The procedure stops at the time moment ϑ .

Construction of a family D_h is based on Theorem 1 formulated below. Let us fix a functional $\Lambda^0(\cdot, \cdot)$ on $W_*(T) \times W_*(T)$.

Definition 1 [7, 15] A family D_h , $h \in (0, 1)$, (8) of positional algorithms of modeling is said to be Λ^0 -stable if there exist functions $k_1(\cdot)$, $k_2(\cdot)$, $k_3(\cdot)$: $[0, \infty) \rightarrow [0, \infty)$ such that $k_1(h) \rightarrow 1$, $k_2(h) \rightarrow 0$, $k_3(h) \rightarrow 0$ as $h \rightarrow 0$, and for every observation record $\xi^h(\cdot) \in \Xi(x_r(\cdot), h)$, it holds that

$$|v^h(\cdot)|_{L_2(T; U)} \leq k_1(h)|u_*(\cdot; x_r(\cdot))|_{L_2(T; U)} + k_2(h), \quad (9)$$

$$\Lambda^0(x_r(\cdot), w^h(\cdot)) \leq k_3(h), \quad (10)$$

where $v^h(\cdot) = D_h \xi^h(\cdot)$ and $w^h(\cdot)$ is the model motion generated by D_h under the observation record $\xi^h(\cdot)$.

Then the following theorem is true.

Theorem 1 [15] *Let a family D_h of positional algorithms of modeling be a) Λ^0 -stable, and b) for every $h_k > 0$, $h_k \rightarrow 0+$ as $k \rightarrow +\infty$, $\xi^{h_k}(\cdot) \in \Xi(x_r(\cdot), h_k)$, $w^{h_k}(\cdot) = w^{h_k}(\cdot; t_0, w^{h_k}(t_0), v^{h_k}(\cdot))$, $v^{h_k}(\cdot) = D_{h_k} \xi^{h_k}(\cdot)$, the limit correlations $v^{h_k}(\cdot) \rightarrow v(\cdot)$ weakly in $L_2(T; U)$, $\Lambda^0(x_r(\cdot), w^{h_k}(\cdot)) \rightarrow 0$ imply that $v(\cdot) \in U(x_r(\cdot))$. Then the family D_h , $h \in (0, 1)$, is regularizing.*

2. Case $u_*(\cdot; x_r(\cdot)) \in L_\infty(T; U)$

First we consider a case where a control $u_*(\cdot; x_r(\cdot))$ is a bounded function, i.e., $u_*(\cdot; x_r(\cdot)) \in L_\infty(T; U)$. Let a number $a > 0$ such that $x_0 \in X_0 = \{x \in D(\varphi) : |x|_H^2 + \varphi_\omega(x) \leq a < +\infty\}$ be chosen.

We take a family of partitions (Δ_h) of the interval T , and functions $\alpha(h):R^+ \rightarrow R^+$ and $d(h):R^+ \rightarrow R^+$ satisfying the following conditions:

$$\begin{aligned} \alpha(h) &\rightarrow 0, \quad \delta(h) \rightarrow 0, \quad d(h) \rightarrow +\infty, \\ d(h)(h + \delta(h))/\alpha(h) &\rightarrow 0, \quad \alpha(h)d^2(h) \rightarrow 0 \text{ as } h \rightarrow 0+. \end{aligned}$$

This condition holds, for example, if $\delta(h) \leq ch$, $\alpha(h) = c_1 h^\gamma$, $\gamma = 0, 5 + \gamma_0$, $\gamma_0 \in (0; 0, 25)$, $d(h) = c_2 h^{-1/4}$, c, c_1, c_2 are positive constants. The family (\mathcal{W}_h) of t_0 -algorithms is defined by rule (5), where

$$w_0^h \in B(\xi_0^h) = \{x \in X_0 : |\xi_0^h - x|_H \leq 2h\}. \quad (11)$$

From (3) and inclusion $x_0 \in X_0$, it follows $B(\xi_0^h) \neq \emptyset$. In its turn, the model control law $S_h = (\Delta_h, \mathcal{U}_h)$ is defined following rule (6), (7), where we assume

$$\mathcal{U}_h(\tau_i, \xi_i^h, w^h(\tau_i)) = \arg \min \{l(\alpha, v, s_i) : v \in S(d(h))\}, \quad (12)$$

$$\begin{aligned} l(\alpha, v, s_i) &= 2(s_i, Bv) + \alpha(h)|v|_U^2, \quad s_i = w^h(\tau_i) - \xi_i^h, \\ S(d(h)) &= \{u \in U : |u|_U \leq d(h)\}. \end{aligned} \quad (13)$$

For a fixed h (and, consequently, a fixed family $\Delta_h = \{\tau_i\}_{i=0}^{m_h}$ with diameter $\delta(h)$) model M is given by the mapping associated with each triple $(w_0^h, \xi^h(\cdot), v^h(\cdot))$, $w_0^h \in X_0$, $\xi^h(\cdot) \in \Xi(x(\cdot), h)$, $v^h(\cdot) \in L_2(T; U)$ with a function $w^h(\cdot) = w^h(\cdot; t_0, w_0^h, \xi^h(\cdot), v^h(\cdot)) \in W_*(T)$ which is a unique solution of variational inequality (4), where

$$B_t(\xi^h(\cdot), w^h(\cdot))v^h(t) = Bv^h(t) - \omega^*(w^h(\tau_i) - \xi_i^h) \quad (14)$$

for $t \in \delta_i = [\tau_i, \tau_{i+1})$, $\tau_i = \tau_{h,i}$, $\xi_i^h = \xi^h(\tau_i)$, $\omega^* = 0$, if $\omega \leq 0$, $\omega^* = \omega + \varepsilon_0$, otherwise ($\varepsilon_0 \geq 0$). Let

$$\Lambda^0(x(\cdot), w^h(\cdot)) = |x(\cdot) - w^h(\cdot)|_{C(T; H)}^2 + 2c|x(\cdot) - w^h(\cdot)|_{L_2(T; V)}^2. \quad (15)$$

Theorem 2 *The family of positional modeling algorithms D_h (8) of the form (4)–(7), (11)–(14) satisfies conditions of Theorem 1 and is regularizing.*

Before proving the theorem, we formulate auxiliary statements. Let $\omega_* = \omega$, if $\omega > 0$, $\omega_* = 0$ otherwise. Introduce a mapping $l(\cdot) : W_*(T) \rightarrow R^+$,

$$l(y(\cdot)) = |y(\cdot)|_{C(T; H)} + |y_t(\cdot)|_{L_2(T; H)} + |y(\cdot)|_{L_2(T; V)}.$$

The following lemmas are true.

Lemma 1 *There exists a number $K_* = K_*(\omega_*, c, |B|_{L(T; U)})$ such that for any $x_0 \in D(\varphi)$, $u(\cdot) \in U_T$, $x(\cdot) = x(\cdot; t_0, x_0, u(\cdot))$ the inequality*

$l(x(\cdot)) \leq K_*(1 + \omega_*|x_0|_H + \varphi_\omega^{1/2}(x_0) + |u(\cdot)|_{L_2(T;U)})$
holds.

Lemma 2 *There exists a number $K^* = K^*(\omega_*, c, |B|_{\mathcal{L}(T;U)}, \varepsilon_0)$ such that for any $w_0^h \in D(\varphi)$, $v^h(\cdot) \in U_T$, $\xi^h(\cdot) \in \Xi(x_r(\cdot), h)$, $w^h(\cdot) = w^h(\cdot; t_0, w_0^h, \xi^h(\cdot), v^h(\cdot))$, the inequality*

$$l(w^h(\cdot)) \leq K^*(1 + \omega_*|w_0^h|_H + \varphi_\omega^{1/2}(w_0^h) + |v^h(\cdot)|_{L_2(T;U)})$$

holds.

These lemmas are proved by the standard scheme [1, 2].

Let us emphasize that constants K_* and K^* do not depend on w_0^h , $\xi^h(\cdot)$, $u(\cdot)$ or $v^h(\cdot)$.

Proof of Theorem 2. Let us show that the family D_h (4)–(7), (11)–(14) is Λ^0 -stable. Let $\xi^h(\cdot) \in \Xi(x_r(\cdot), h)$, $v^h(\cdot) = D_h \xi^h(\cdot)$ and $w^h(\cdot)$ is the model motion (4) generated by D_h under the observation record $\xi^h(\cdot)$. Consider the value

$$\varepsilon_h(t) = |\mu^h(t)|_H^2 + 2c|\mu^h(\cdot)|_{L_2([t_0, t]; V)}^2 + \alpha(h) \int_{t_0}^t \{|v^h(\tau)|_U^2 - |u_*(\tau)|_U^2\} d\tau.$$

Here and below $\mu^h(\cdot) = x_r(\cdot) - w^h(\cdot)$. Setting in (1) and (4) $z = w^h(t)$ and $z = x_r(t)$, respectively, and summing the corresponding inequalities, we obtain

$$(\mu_t^h(t), \mu^h(t)) + \langle A\mu^h(t), \mu^h(t) \rangle \leq (B(u_*(t) - v^h(t)), \mu^h(t)) - \\ \omega^*(w^h(\tau_i) - \xi_i^h, \mu^h(t)) \quad \text{for a.a. } t \in \delta_i = [\tau_i, \tau_{i+1}).$$

($\tau_i = \tau_{h,i}$, $\xi_i^h = \xi^h(\tau_{h,i})$). Hence, taking into account (2) we derive a.e. in δ_i

$$\frac{1}{2}d|\mu^h(t)|_H^2/dt + c|\mu^h(t)|_V^2 - \omega|\mu^h(t)|_H^2 \leq \\ (B(u_*(t) - v^h(t)), \mu^h(t)) + \omega^*(w^h(\tau_i) - \xi_i^h, \mu^h(t)). \quad (16)$$

It is easy to see that for $t \in \delta_i$

$$\leq (h + \int_{\tau_i}^t \{|w_\tau^h(\tau)|_H + |x_{r\tau}(\tau)|_H\} d\tau) |\mu^h(t)|_H - |\mu^h(t)|_H^2. \quad (17)$$

In this case, for a.a. $t \in \delta_i$

$$\frac{1}{2}d|\mu^h(t)|_H^2/dt + c|\mu^h(t)|_V^2 \leq (B(u_*(t) - v^h(t)), \xi_i^h - w^h(\tau_i)) + \rho_i(t, h, \delta), \quad (18)$$

where $k_h = \omega^* + |B|(Q + d(h))$, $Q = |u_*(\cdot)|_{L_\infty(T;U)}$,

$$\rho_i(t, h, \delta) = k_h(h + \int_{\tau_i}^t \{|w_\tau^h(\tau)|_H + |x_{r\tau}(\tau)|_H\} d\tau).$$

Let $h_* > 0$ such that $d(h) > Q$ for $h \in (0, h_*)$. Hence, using the rule of definition of a control $v^h(\cdot)$, we obtain

$$\varepsilon_h(t) \leq \varepsilon(\tau_i) + 2(t - \tau_i)\rho_i(t, h, \delta), \quad t \in \delta_i.$$

Consequently, by Lemmas 1 and 2, the inequality

$$\varepsilon_h(t) \leq \varepsilon_h(t_0) + 2k_h(h(\vartheta - t_0) + \delta(h)) \int_{t_0}^t \{ |w_\tau^h(\tau)|_H + |x_{r\tau}(\tau)|_H \} d\tau \leq k_1(1 + d(h))(h + \delta(h)) \quad t \in T$$

is valid. We deduce

$$\Lambda^0(x_r(\cdot), w^h(\cdot)) \leq k_1(1 + d(h))(h + \delta(h)) + \alpha(h)k_2(1 + d^2(h)), \quad (19)$$

$$|v^h(\cdot)|_{L_2(T;U)} \leq |u_*(\cdot)|_{L_2(T;U)} + k_1(\alpha^{-1}(h)(1 + d(h))(h + \delta(h)))^{1/2}. \quad (20)$$

Here, the constant k_1 does not depend on h and can be written explicitly. Thus, the family D_h is Λ^0 -stable. It is easy to see that the condition b) of Theorem 1 holds, if

$$y(\cdot) = x_r(\cdot), \quad (21)$$

where $y(\cdot) = x(\cdot; t_0, x_0, v(\cdot))$. Prove (21). Note, that

$$|x_r(\cdot) - w^{h_k}(\cdot)|_{C(T;H)} \rightarrow 0 \quad \text{as } k \rightarrow \infty, \quad (22)$$

if $h_k \rightarrow 0$. Here $w^{h_k}(\cdot) = w(\cdot; t_0, w^{h_k}(t_0), \xi^{h_k}(\cdot), v^{h_k}(\cdot))$, $v^{h_k}(\cdot) = D_{h_k}(\xi^{h_k}(\cdot))$. The inequality

$$\begin{aligned} & |y(t) - w^{h_k}(t)|_H^2 + 2 \int_{t_0}^t \{ c|y(\tau) - w^{h_k}(\tau)|_V^2 - \omega|y(\tau) - w^{h_k}(\tau)|_H^2 \} d\tau \leq \\ & \quad 2 \int_{t_0}^t (B(v(\tau) - v^{h_k}(\tau)), y(\tau) - w^{h_k}(\tau))_H d\tau + \\ & \quad |\omega^*| \max_{i \in [0:m_{h_k}]} |\xi_i^{h_k} - w^{h_k}(\tau_i)|_H \int_{t_0}^t |y(\tau) - w^{h_k}(\tau)|_H d\tau \end{aligned}$$

is true. Hence, taking into account (22), inclusion $\xi^{h_k}(\cdot) \in \Xi(x_r(\cdot), h_k)$, convergence h_k to 0, and weak convergence of controls $v^{h_k}(\cdot)$ to $v(\cdot)$ in $L_2(T;U)$ as $k \rightarrow \infty$, in virtue of Gronwall inequality, we obtain

$$|y(\cdot) - w^{h_k}(\cdot)|_{C(T;H)} \rightarrow 0 \quad \text{as } k \rightarrow \infty.$$

Therefore, relation (21) holds. The theorem follows from Theorem 1.

Let the following condition be fulfilled

Condition 2 The function φ is differentiable and operator $Cx = \text{grad}\varphi(x) : V \rightarrow V^*$ is Lipschitz.

Then the estimate of convergence rate of the algorithm is true.

Theorem 3 Let $U = V$, B be the operator of canonical embedding V in H and $u_*(\cdot) = u_*(\cdot; x_r(\cdot))$ be a function with bounded variation. Then the estimate

$$|u_*(\cdot) - v^h(\cdot)|_{L_2(T;H)} \leq K \{ [d(h)(h + \delta(h)) + \alpha(h)d^2(h)]^{1/2} + \alpha^{-1}(h)d(h)(h + \delta(h)) \}$$

holds.

The theorem is proved analogously to [15].

3. Case $u_*(\cdot; x_r(\cdot)) \in L_2(T; U)$

Consider the case when $u_*(\cdot; x_r(\cdot))$ is a norm-square integrable function, i.e., $u_*(\cdot; x_r(\cdot)) \in L_2(T; U)$. As the model M we take the variational inequality (4) with the family of operators $B_t(\cdot; \cdot) : U \rightarrow H$ of the form (14). The family (\mathcal{W}_h) of t_0 -algorithms is defined as mentioned above, i.e., by rule (5), (11), the model control law $\mathcal{S}_h = (\Delta_h, \mathcal{U}_h)$ being defined by the rule (6), (7), where we assume

$$\mathcal{U}_h(\tau_i, \xi_i^h, w^h(\tau_i)) = \arg \min \{l(\alpha, v, s_i) : v \in U\} = -\alpha^{-1} B' s_i, \quad (23)$$

$s_i = w^h(\tau_i) - \xi_i^h$. Let a family of partitions (Δ_h) of the interval T , and function $\alpha(h) : R^+ \rightarrow R^+$ satisfying the following conditions:

$$\begin{aligned} h\delta^{-1}(h) &\leq C, \quad \delta(h)\alpha^{-2}(h) \leq C, \\ \alpha(h) &\rightarrow 0, \quad \delta(h) \rightarrow 0, \quad (h + \delta(h))/\alpha(h) \rightarrow 0 \quad \text{as } h \rightarrow 0+. \end{aligned} \quad (24)$$

Here $C = \text{const} > 0$ that does not depend on h .

Theorem 4 *The family of positional modeling algorithms D_h (8) of the form (4)–(7), (11), (14), (23) satisfies conditions of Theorem 1 and is regularizing.*

Proof of theorem is carried out by the scheme of that for Theorem 1. Taking into account (2) we derive a.e. in δ_i (16). It is also easy to see that for $t \in \delta_i$ inequality (17) holds along with

$$\begin{aligned} (B(u_*(t) - v^h(t), \mu^h(t)) &\leq (B(u_*(t) - v^h(t)), \xi_i^h - w^h(\tau_i)) + \\ &+ |B|\{|u_*(t)| + |v^h(t)|\}(h + \int_{\tau_i}^t g^h(\tau) d\tau), \end{aligned}$$

In this case, for a.a. $t \in \delta_i$ estimate (18) holds, provided

$$\begin{aligned} \rho_i(t, h, \delta) &= \{\omega^* + |B|(|u_*(t)|_U + |v^h(t)|_U)\}(h + \int_{\tau_i}^t g^h(\tau) d\tau), \\ g^h(\tau) &= |w_\tau^h(\tau)|_H + |x_{r\tau}(\tau)|_H. \end{aligned}$$

Hence, using the rule of definition of a control $v^h(\cdot)$ (23), we obtain

$$\begin{aligned} \varepsilon_h(t) &\leq \varepsilon_h(\tau_i) + \omega^*\delta h + \delta\omega^* \int_{\tau_i}^t g^h(\tau) d\tau + h^2|B|^2 + \\ 3\delta \int_{\tau_i}^t F^h(\tau) d\tau &+ 2\delta^2(1 + |B|^2) \int_{\tau_i}^t \{|w_\tau^h(\tau)|_H^2 + |x_{r\tau}(\tau)|_H^2\} d\tau, \end{aligned} \quad (25)$$

where $F^h(\tau) = |u_*(\tau)|_U^2 + |v^h(\tau)|_U^2$. Summing right and left hand of the inequality (25) over i and taking into account Lemmas 1 and 2, the

inclusion $u_*(\cdot) \in L_2(T; U)$, we deduce

$$\begin{aligned} \varepsilon_h(t) &\leq \varepsilon_h(t_0) + b_1 h(1 + h/\delta) + b_2 \delta + b_3 \delta \int_{t_0}^t F^h(\tau) d\tau \leq \\ &\leq \varepsilon_h(t_0) + b_1 h(1 + h/\delta) + b_4 \delta + b_3 \delta^2 \sum_{j=0}^{i(t)} |v_j^h|^2, \end{aligned} \quad (26)$$

where the symbol $i(t)$ denotes an integer part of number t . Besides, by the rule of definition of v_i^h (see (23)), we have

$$|v_i^h|_U^2 \leq 2|B|^2(\mu_i^h + h^2)\alpha^{-2}, \quad (27)$$

where $\mu_i^h = \mu^h(\tau_i) = x_r(\tau_i) - w^h(\tau_i)$. From (26), (27) and inequalities $\varepsilon_h(t_0) \leq 4h^2$, $h\delta^{-1}(h) \leq C$, follows the estimation

$$\begin{aligned} \mu_i^h &\leq \varepsilon_h(t_0) + b_1 h(1 + h/\delta) + b_3 \delta + \alpha|u_*(\cdot)|_{L_2(T; U)}^2 + \\ &+ b_2 \delta^2 \sum_{j=0}^{i-1} 2|B|^2(\mu_j^h + h^2)\alpha^{-2} \leq b_5(h + \delta + \alpha) + b_6 \delta^2 \alpha^{-2} \sum_{j=0}^{i-1} \mu_j^h. \end{aligned}$$

Taking into account Gronwoll inequality, and inequality $\delta(h)\alpha^{-2}(h) \leq C$, we deduce

$$\mu_i^h \leq b_5(h + \delta + \alpha) \exp\{b_6(\vartheta - t_0)\delta/\alpha^2\} \leq b_7(h + \delta + \alpha). \quad (28)$$

Summing left hand of the inequality (27) over i , we obtain from (28)

$$\delta^2 \sum_{j=0}^{m_h-1} |v_j^h|^2 \leq 2|B|^2 \sum_{j=0}^{m_h-1} (\mu_j^h + h^2)\alpha^{-2} \leq b_8 \delta \alpha^{-2}(\alpha + h + \delta). \quad (29)$$

Due to (24) without commonness loss, suppose $\delta(h)\alpha^{-1}(h) \leq 1$. Therefore, from (26), (29) we derive the estimation

$$\varepsilon_h(t) \leq b_7(h + \delta + \delta^2\alpha^{-2} + h\delta\alpha^{-2}) \leq b_8(h + \delta). \quad (30)$$

Therefore $|v^h(\cdot)|_{L_2(T; U)}^2 \leq |u_*(\cdot)|_{L_2(T; U)}^2 + b_8(h + \delta)\alpha^{-1}$. This means that the inequality (9) takes place if we have $k_1(h) = 1$, $k_2(h) = (b_8(h + \delta))\alpha^{-1}(h)^{1/2}$. From (30) we deduce (10), where functional $\Lambda^0(\cdot, \cdot)$ is defined by (15) and $k_3(h) = b_9(h + \delta(h) + \alpha(h))$. The theorem is proved.

The next theorem is proved analogously to Theorem 3:

Theorem 5 *Let conditions of Theorem 3 hold. Then the estimate*

$$|u_*(\cdot) - v^h(\cdot)|_{L_2(T; H)} \leq K\{(h + \delta(h))^{1/2} + (h + \delta(h))\alpha^{-1}(h)\}$$

is true.

References

- [1] Brezis, H. (1973). *Operateurs maximaux monotones et semigroupes de contractions dans les espaces de Hilbert*. Amsterdam–London–New York.
- [2] Barbu, V. (1984). *Optimal control of variational inequalities*. Pitman Advanced Publishing Program, London.
- [3] Duvaut, G. and Lions, J.-L. (1972). *Les inéquations en mécanique et en physique*. Dunod, Paris.
- [4] Glowinski, R., Lions, J.-L. and Trémolières, R. (1980). *Numerical analysis of variational inequalities*. North-Holland, Amsterdam.
- [5] Kryazhimskii, A. V. and Osipov, Yu. S. (1983). Modelling of a control in a dynamic system. *Engineering Cybernetics*, 21:38–47. in Russian.
- [6] Kryazhimskii, A. V., Maksimov, V. I. and Osipov, Yu. S. (1983). On positional simulation in dynamic systems. *J. Appl. Math. Mech.*, 47:709–714.
- [7] Osipov, Yu. S., Kryazhimskii, A. V. and Maksimov, V. I. (1991). *Dynamic regularization problems for distributed parameter systems*. Sverdlovsk, Russia. in Russian.
- [8] Kryazhimskii, A. V. and Osipov, Yu. S. (1995). *Inverse problems for ordinary differential equations: dynamical solutions*. Gordon and Breach, London.
- [9] Maksimov, V. I. (1988). Dynamical modeling of unknown disturbances in parabolic variational inequalities. *Prikl. Mat. Mech.*, 52:743–750. in Russian.
- [10] Maksimov, V. I. (1992). Inverse problems for variational inequalities. *Internat. ser. of Numer. Math.*, 107:275–286. Birkhäuser Verlag, Basel.
- [11] Maksimov, V. I. (1990). *Stable reconstruction of unknown disturbances in parabolic variational inequalities*. In: Problems of Optimization and stability in Control Systems. UB AS USSR, Sverdlovsk. 74–86. in Russian.
- [12] Pandolfi, L. and Maksimov, V.I. (2000). Dynamical reconstruction of unbounded controls in nonlinear dynamical systems. Proceedings CD of the Fourteenth International Symposium of Mathematical Theory of Networks and Systems MTNS , Perpignan, France, June 19–23, 2000.
- [13] Maksimov, V. (1998). Dynamical inverse problems for parabolic variational inequalities. *An. St. Univ. Ovidius Constanta*, 6:97-110.
- [14] Osipov, Yu. S., Kryazhimskii, A. V. and Maksimov, V. I. (2000). Dynamical inverse problems for parabolic systems. *Differential Equations*, 36:579–597. in Russian.
- [15] Maksimov, V. I. (2000). *The problems of dynamical reconstruction of inputs of infinite-dimensional systems*. Ekaterinburg. in Russian.
- [16] Kappel, F., Pandolfi, L., Maksimov, V. and Blizorukova, M. (2000). Problems of dynamical identification in nonlinear differential systems. *Proceedings of the Polish-German symposium on "Science Research Education" (SRE'2000)*, 51–56.
- [17] Krasovskii, N. N. and Subbotin, A. I. (1988). *Game-theoretical control problems*. Springer.

FLOW-INVARIANCE PROPERTIES FOR A CLASS OF DISCRETE-TIME NONLINEAR UNCERTAIN SYSTEMS

Laurentiu Marinovici and Octavian Pastravanu

Department of Automatic Control and Industrial Informatics

Technical University "Gh. Asachi" of Iasi

Bld. Mangeron 53A, 6600 Iasi, Romania

Phone: +40-32-230751, Fax: +40-32-214290

lmarinovici@ac.tuiasi.ro

opastrav@ac.tuiasi.ro

Abstract The family of time-dependent rectangular sets, flow invariant with respect to a class of discrete-time nonlinear uncertain systems, is studied. The invariance results are further used for dealing with a special type of asymptotic stability, called componentwise asymptotic stability (CWAS), which can be characterized by difference inequalities. The particularisation of the CWAS conditions for exponential type time-dependence yields the stronger property of componentwise exponential asymptotic stability (CWEAS), that is proven equivalent to some algebraic inequalities.

Keywords: Flow-invariant sets, discrete-time systems, nonlinear systems, uncertain systems.

AMS Subject Classification: 93C10, 93D20, 93C55, 93C41

1. Introduction

Using the powerful tool offered by the flow-invariance theory, a componentwise refinement is envisaged in approaching the dynamics of a class of discrete-time nonlinear systems with uncertainties. Such a refinement, by individually monitoring each state variable, is able to reveal some important properties of the trajectories around the equilibrium points, which remain hidden within the standard framework of stability analysis. The usage of flow-invariance concepts for a componentwise exploration of the dynamical system behavior appeared in mid eighties

in Pavel's and Voicu's works ([4]), ([5]), ([6]), ([7]), focusing exclusively on the continuous-time case. The discrete-time case was addressed later in papers, such as ([1]) for 1-D and 2-D linear systems and ([2]) for linear systems with uncertainties. All the above cited works pointed out that the componentwise investigation naturally leads to particular types of asymptotic stability, namely the *componentwise asymptotic stability* (CWAS) and *componentwise exponential asymptotic stability* (CWEAS), as initially introduced in ([5]), ([6]).

The class of nonlinear uncertain systems studied in the current paper may be roughly regarded as a discrete-time counterpart of the continuous-time dynamics discussed in ([3]). In spite of the similarities occurring at the first glance between the mathematical description of the continuous-time and discrete-time cases, this deeper insight emphasizes noticeable differences, which are commented throughout the text. Our material is structured as follows. Section 2 deals with the existence of *time-dependent rectangular sets* (TDRSs), which are *flow-invariant* (FI) with respect to a given *discrete-time nonlinear uncertain system* (DTNUS). Sections 3 and 4, by using the FI results, analyse CWAS, and respectively, CWEAS of the equilibrium point (EP) $\{0\}$. Some concluding remarks, including comparisons with the continuous-time case in ([3]) are formulated in Section 5. Since the complete proofs of most results are laborious this version of the text limits them to some basic explanation.

2. The Family of TDRSs with FI Properties

Consider the class of *discrete-time nonlinear uncertain systems* (DTNUSs) defined as:

$$\mathbf{x}(t+1) = \mathbf{f}(\mathbf{x}(t)), \mathbf{x} \in \mathbb{R}^n, \mathbf{x}(t_0) = \mathbf{x}_0, t \in \mathbb{Z}_+, \quad (1)$$

$$f_i(\mathbf{x}(t)) = \sum_{j=1}^n a_{ij} x_j^{p_{ij}}(t), p_{ij} \in \mathbb{N}, i = \overline{1, n},$$

with *interval-type coefficients*:

$$a_{ij}^- \leq a_{ij} \leq a_{ij}^+, i, j = \overline{1, n}. \quad (2)$$

Also consider the n -valued vector function $\mathbf{v}(t) : \mathbb{Z}_+ \rightarrow \mathbb{R}^n$ with positive components $v_i(t) > 0, i = \overline{1, n}$. Using these $v_i(t) > 0, i = \overline{1, n}$ define the *time-dependent rectangular set* (TDRS).

$$\mathbf{H}_v(t) = [-v_1(t), v_1(t)] \times \dots \times [-v_n(t), v_n(t)], \quad (3)$$

where $[,] \times [,]$ denotes the Cartesian product.

Let us explore the free response of DTNUS (1) along the lines of the componentwise constrained evolution of the state trajectories induced by the concept of *flow-invariance* (FI).

Definition 1 TDRS (3) is called *flow-invariant* (FI) with respect to (w.r.t.) DTNUS (1) if for any initial condition $\mathbf{x}(t_0) = \mathbf{x}_0 \in \mathbf{H}_v(t_0)$, the corresponding state trajectory $\mathbf{x}(t) = \mathbf{x}(t; t_0, \mathbf{x}_0)$ remains (for all possible values resulting from the interval-type coefficients) inside $\mathbf{H}_v(t)$, for $t \in [t_0, \infty)$, i.e.

$$\begin{aligned} \forall \mathbf{x}(t_0) = \mathbf{x}_0 \in \mathbf{H}_v(t_0) : \\ \mathbf{x}(t) = \mathbf{x}(t; t_0, \mathbf{x}_0) \in \mathbf{H}_v(t), t \in [t_0, \infty). \end{aligned} \quad (4)$$

Theorem 1 TDRS (3) is FI w.r.t. DTNUS (1) iff the following vector inequalities hold for $t \in [t_0, \infty)$:

$$\mathbf{v}(t+1) \geq \bar{\mathbf{g}}(\mathbf{v}(t)); \bar{g}_i(\mathbf{v}(t)) = \sum_{j=1}^n \bar{c}_{ij} v_j^{p_{ij}}(t), i = \overline{1, n}; \quad (5-a)$$

$$\mathbf{v}(t+1) \geq \tilde{\mathbf{g}}(\mathbf{v}(t)); \tilde{g}_i(\mathbf{v}(t)) = \sum_{j=1}^n \tilde{c}_{ij} v_j^{p_{ij}}(t), i = \overline{1, n}, \quad (5-b)$$

where \bar{c}_{ij} , \tilde{c}_{ij} have unique values, derived from the interval-type coefficients a_{ij} of DTNUS (1) as follows:

$$\bar{c}_{ij} = \begin{cases} \max \left\{ \left| a_{ij}^- \right|, \left| a_{ij}^+ \right| \right\}, & \text{if } p_{ij} \text{ odd} \\ \max \left\{ 0, a_{ij}^+ \right\}, & \text{if } p_{ij} \text{ even} \end{cases}, \quad (6-a)$$

$$\tilde{c}_{ij} = \begin{cases} \max \left\{ \left| a_{ij}^- \right|, \left| a_{ij}^+ \right| \right\}, & \text{if } p_{ij} \text{ odd} \\ \max \left\{ 0, -a_{ij}^- \right\}, & \text{if } p_{ij} \text{ even} \end{cases}. \quad (6-b)$$

Proof. It results by expressing the bounds of the right-hand side of (1) in terms of a_{ij}^- , a_{ij}^+ , $i, j = \overline{1, n}$ in (2) and $v_i(t)$, $i = \overline{1, n}$, in (3). ■

Theorem 2 There exist TDRSs (3) which are FI w.r.t. DTNUS (1) iff there exist positive solutions (PSSs) $y_i(t) > 0$, $i = \overline{1, n}$ for the following difference inequalities (DIs):

$$\mathbf{y}(t+1) \geq \bar{\mathbf{g}}(\mathbf{y}(t)) \quad (7-a)$$

$$\mathbf{y}(t+1) \geq \tilde{\mathbf{g}}(\mathbf{y}(t)). \quad (7-b)$$

Proof. On the basis of Definition 1 and Theorem 1, any positive solution $y_i(t) > 0, i = \overline{1, n}$ of the inequality system (7-a), (7-b) defines a flow invariant TDRS (3). ■

Remark 1 The algorithm for constructing the coefficients (6-a) and (6-b) of DIs (7-a) and (7-b) differs from its continuous-time counterpart studied in ([3]), yielding only positive values for these coefficients as also occurring for discrete-time linear systems, e.g. ([1]), ([2]).

Theorem 3 *There exist TDRSs (3) which are FI w.r.t. DTNUS (1) iff there exist PSs $y_i(t) > 0, i = \overline{1, n}$ for the following DI:*

$$\mathbf{y}(t+1) \geq \mathbf{g}(\mathbf{y}(t)); \mathbf{g}: \mathbb{R}^n \rightarrow \mathbb{R}^n, g_i(\mathbf{y}) = \max_{\mathbf{y} \in \mathbb{R}^n} \{\bar{g}_i(\mathbf{y}), \tilde{g}_i(\mathbf{y})\}, i = \overline{1, n}. \quad (8)$$

Proof. Inequality (8) replaces the two inequalities (7-a) and (7-b) from Theorem 2 in an equivalent manner. ■

In order to investigate the family of TDRSs, which are FI w.r.t. a given DTNUS, we will first focus on some relevant characteristics of the PSs of the inequality (8), since Theorem 3 emphasizes a bijective link between the two types of mathematical objects. We start with the qualitative exploration of the solution of the following *difference equation* (DE):

$$\mathbf{z}(t+1) = \mathbf{g}(\mathbf{z}(t)), t \in \mathbb{Z}_+ \quad (9)$$

which is obtained from DI (8) by replacing " \geq " with " $=$ ".

Theorem 4 *For any $t_0 \in \mathbb{Z}_+$ and any positive initial condition $\mathbf{z}(t_0) = \mathbf{z}_0 > 0$, the unique solution $\mathbf{z}(t) = \mathbf{z}(t; t_0, \mathbf{z}_0)$ of DE (9) remains positive for any $t \in [t_0, \infty)$.*

Proof. According to the definition of \bar{g}_i and \tilde{g}_i in (5-a) and (5-b), all the coefficients $\bar{c}_{ij}, \tilde{c}_{ij}, i, j = \overline{1, n}$ are non-negative, and thus, for any $t_0 \in \mathbb{Z}_+$ and any positive initial condition $\mathbf{z}(t_0) = \mathbf{z}_0 > 0$, since $g_i(\mathbf{z}) = \max_{\mathbf{z} \in \mathbb{R}^n} \{\bar{g}_i(\mathbf{z}), \tilde{g}_i(\mathbf{z})\}$, the solution $\mathbf{z}(t) = \mathbf{z}(t; t_0, \mathbf{z}_0)$ of DE (9) remains positive for any $t \in [t_0, \infty)$. ■

One can easily see that Theorem 4 guarantees the existence of PSs for DI (8) in the particular case when " \geq " is replaced by " $=$ ". However DI (8) might have PSs, which do not satisfy DE (9) and therefore, a connection between the PSs of DI (8) and the PSs of DE (9) must be further established.

Theorem 5 *Let $\mathbf{y}(t) > 0$ be an arbitrary PS of DI (8), with $t \in [t_0, \infty)$, $t_0 \in \mathbb{Z}_+$. Denote by $\mathbf{z}(t)$ an arbitrary PS of DE (9), corresponding to*

an initial condition $\mathbf{z}(t_0)$, which satisfies the componentwise inequality $0 < \mathbf{z}(t_0) \leq \mathbf{y}(t_0)$. Denote by $\mathbf{z}^*(t)$ the unique PS of DE (9) corresponding to the initial condition taken by $\mathbf{y}(t)$, i.e. $\mathbf{z}^*(t_0) \equiv \mathbf{y}(t_0)$. For $t \in [t_0, \infty)$ the following inequalities hold: $0 < \mathbf{z}(t) \leq \mathbf{z}^*(t) \leq \mathbf{y}(t)$.

Proof. The fulfillment of the inequality $0 < \mathbf{z}(t)$ for $0 < \mathbf{z}(t_0)$ is guaranteed by Theorem 4.

Since $0 < \mathbf{z}(t_0) \leq \mathbf{z}^*(t_0) \equiv \mathbf{y}(t_0)$, it results that: $0 < g_i(\mathbf{z}(t_0)) \leq g_i(\mathbf{z}^*(t_0)) = g_i(\mathbf{y}(t_0))$, $i = \overline{1, n}$, and so $0 < \mathbf{z}(t_0 + 1) \leq \mathbf{z}^*(t_0 + 1) \leq \mathbf{y}(t_0 + 1)$.

By mathematical induction, one can easily see that the inequalities at time $(t - 1), t \in \mathbb{Z}_+$ imply the truthfulness of the same inequalities at time $(t), t \in \mathbb{Z}_+$. ■

Remark 2 Unlike the continuous-time case studied in ([3]), where the FI property might be restricted to a finite-time horizon $[t_0, T)$, the existence of FI w.r.t. DTNUS (1) is guaranteed for an infinite-time horizon $[t_0, \infty)$.

Theorem 6 If $\mathbf{H}_y(t), \mathbf{H}_{z^*}(t)$ and $\mathbf{H}_z(t)$ denote three TDRSs, FI w.r.t. DTNUS (1), generated by the following three types of PSs of DI (8): $\mathbf{y}(t)$ - arbitrary PS of DI (8); $\mathbf{z}^*(t)$ - unique PS of DE (9), with $\mathbf{z}^*(t_0) \equiv \mathbf{y}(t_0)$; $\mathbf{z}(t)$ - arbitrary PS of DE (9), with $\mathbf{z}(t_0) \leq \mathbf{y}(t_0)$, then:

$$\mathbf{H}_z(t) \subseteq \mathbf{H}_{z^*}(t) \subseteq \mathbf{H}_y(t), \forall t \in [t_0, \infty).$$

Proof. It results from Theorems 5 and 3 and Definition 1. ■

Given a TDRS, which is FI w.r.t. DTNUS (1), there can be formulated a condition for the existence of other TDRSs, strictly included in the former one, which are also FI w.r.t. DTNUS (1).

Theorem 7 Denote by $\mathbf{H}_y(t)$ a TDRS, FI w.r.t. DTNUS (1) for $t \in [t_0, \infty)$. If there exist n functions $\delta_i(t)$, non-decreasing, positive and subunitary $0 < \delta_i(t) < 1, i = \overline{1, n}, t \in \mathbb{Z}_+$, such that:

$$\mathbf{g}(\Delta(t)\mathbf{y}(t)) \leq \Delta(t)\mathbf{g}(\mathbf{y}(t)); \Delta(t) = \text{diag}\{\delta_1, \dots, \delta_n\}, \quad (10)$$

then the TDRS $\mathbf{H}_{\Delta y}(t)$, generated by the vector function $\Delta(t)\mathbf{y}(t)$ is also FI w.r.t. DTNUS (1) and

$$\mathbf{H}_{\Delta y}(t) \subset \mathbf{H}_y(t), t \in [t_0, \infty). \quad (11)$$

Proof. One can show that $\Delta(t)\mathbf{y}(t)$ is a PS of DI (8) and, then, applying Theorem 3 and Definition 1, together with $0 < \delta_i(t) < 1$, completes the proof. ■

Remark 3 Functions $\delta_i(t)$ can be chosen as positive, subunitary constants, case in which the resulting TDRS $\mathbf{H}_{\Delta y}(t)$ is *homotetic* with $\mathbf{H}_y(t)$, taking different transformation factors for each component. When all $\delta_i(t), i = \overline{1, n}$ are equal to the same positive, subunitary constant, the transformation factors are identical for all the components.

The next step in refining the conditions imposed to the TDRS FI w.r.t. DTNUS (1) aims to force the boundedness property by adding a supplementary request for the time-dependence of TDRS, namely to approach $\{0\}$ for $t \rightarrow \infty$. Thus, the concept of FI induces a particular type of *asymptotic stability* (AS) for the *equilibrium point* (EP) $\{0\}$ of DTNUS (1), (stronger than the standard concept based on vector norms in \mathbb{R}^n), which is going to be separately studied in the following section.

3. Componentwise Asymptotic Stability

Let $\mathbf{v}(t) : [t_0, \infty) \rightarrow \mathbb{R}^n$ be a vector function with $v_i(t) > 0, i = \overline{1, n}$ (as considered in Definition 1, which introduces the FI concept) and suppose that $\mathbf{v}(t)$ also has the property:

$$\lim_{t \rightarrow \infty} \mathbf{v}(t) = 0. \quad (12)$$

Definition 2 EP $\{0\}$ of DTNUS (1) is called *componentwise asymptotically stable* w.r.t. $\mathbf{v}(t)$ (CWAS_v) if for any $\mathbf{x}(t_0) = \mathbf{x}_0$ with $|\mathbf{x}(t_0)| \leq \mathbf{v}(t_0)$, the following inequality holds: $|\mathbf{x}(t)| \leq \mathbf{v}(t), t \in [t_0, \infty)$.

Remark 4 Definition 2 can be restated in terms of FI, by taking Definition 1, supplemented with condition (12) for the behavior at the infinity.

On the light of the above remark, the results presented in Theorems 2 and 3 can be immediately transformed to characterize CWAS_v of EP $\{0\}$ of DTNUS (1), yielding the following two theorems.

Theorem 8 EP $\{0\}$ of DTNUS (1) is CWAS_v iff there exist common PSs $\mathbf{v}(t) > 0$ for DIs (7-a) and (7-b), with $\lim_{t \rightarrow \infty} \mathbf{v}(t) = 0$.

Proof. It is a direct consequence of Theorem 2 for the particular case of TDRSs meeting condition (12). ■

Theorem 9 EP $\{0\}$ of DTNUS (1) is CWAS_v iff there exist PSs $\mathbf{v}(t)$ for DI (8), with $\lim_{t \rightarrow \infty} \mathbf{v}(t) = 0$.

Proof. It is a direct consequence of Theorem 3 in the particular case of TDRSs meeting condition (12). ■

The boundedness of TDRSs on $[t_0, \infty)$ together with the requirement (12) introduces some restrictions for the exponents p_{ii} and the interval-type coefficients a_{ii} of DTNUS (1).

Theorem 10 *A necessary condition for EP $\{0\}$ of DTNUS (1) to be CWAS_v is that*

$$\max \{ |a_{ii}^-|, |a_{ii}^+| \} (v_i(t_0))^{p_{ii}-1} < 1, i = \overline{1, n} \quad (13)$$

where $\mathbf{v}(t_0)$ denotes the initial value for the vector function $\mathbf{v}(t)$.

Proof. It results from the fact that the solutions of DEs $v_i(t+1) = \widehat{c}_{ii} v_i^{p_{ii}}(t), \widehat{c}_{ii} = \max \{ \bar{c}_{ii}, \tilde{c}_{ii} \}, i = \overline{1, n}$ should converge to $\{0\}$. ■

Remark 5 Unlike the necessary condition formulated by Theorem 9 in ([3]) requiring p_{ii} - odd and $a_{ii}^+ < 0, i = \overline{1, n}$, the necessary condition given above does not restrict the integer values of the exponents p_{ii} , but it emphasizes the role of the initial values for the vector $\mathbf{v}(t)$ used in Definition 2 of CWAS. The lack of a complete similarity between the continuous-time and discrete-time cases is a consequence of the fact that DTNUS (1) cannot be obtained by the discretization of the continuous-time system considered in ([3]).

Remark 6 The proof of our Theorem 10 shows that by the replacement of " $<$ " by " \leq " in inequalities (13), a necessary condition is obtained for the boundedness of TDRS (3) (without meeting the convergence condition (12) of $\mathbf{v}(t)$ used in Definition 2). This boundedness condition may be regarded as corresponding to the necessary condition in the continuous-time case, given by Theorem 6 in ([3]).

In order to develop a refined interpretation of the result stated in Theorem 9, it is of great interest to resume the qualitative analysis of the solutions of DI (8) and DE (9).

Theorem 11 *Consider an arbitrary PS $\mathbf{y}(t) > 0$ of DI (8), with $t \in [t_0, \infty)$, and the initial condition $\mathbf{y}(t_0)$ satisfying restriction (13) for all $i = \overline{1, n}$. If $\mathbf{z}(t)$ denotes an arbitrary solution of DE (9) corresponding to the initial condition $\mathbf{z}(t_0)$ satisfying: $-\mathbf{y}(t_0) \leq \mathbf{z}(t_0) \leq \mathbf{y}(t_0)$, then the following inequality holds for any $t \in [t_0, \infty)$*

$$-\mathbf{y}(t) \leq \mathbf{z}(t) \leq \mathbf{y}(t). \quad (14)$$

Proof. Starting from the componentwise initial restrictions $-y_i(t_0) \leq z_i(t_0) \leq y_i(t_0), t_0 \in \mathbb{Z}_+$, it is easy to see that $-\mathbf{g}(\mathbf{y}(t_0)) \leq \mathbf{g}(\mathbf{z}(t_0)) \leq$

$\mathbf{g}(\mathbf{y}(t_0))$. This former inequality leads to: $-\mathbf{y}(t_0 + 1) \leq \mathbf{z}(t_0 + 1) \leq \mathbf{y}(t_0 + 1)$. By induction it is proven that inequality (14) holds for any $t \in [t_0, \infty)$. ■

Theorem 11 creates a deeper insight into the topology of the solutions (not only positive) of DE (9) in the vicinity of EP $\{0\}$, which permits revealing the link between condition (12) and the nature of EP $\{0\}$ for DE (9).

Theorem 12 *EP $\{0\}$ of DTNUS (1) is CWAS_v iff EP $\{0\}$ of DE (9) is AS.*

Proof. Sufficiency. One can take a positive initial condition for DE (9) such that, according to Theorem 4, the corresponding solution is positive and also meets condition (12). Then one uses Theorem 9. *Necessity.* DI (8) has PSs meeting condition (12) and applying Theorem 11 one can prove the AS of EP $\{0\}$ of DE (9). ■

Remark 7 It is obvious that the concept of CWAS_v for EP $\{0\}$ of DTNUS (1) is not equivalent to the standard AS. If EP $\{0\}$ of DTNUS (1) is CWAS_v then it is AS, but the converse statement is not true. However, Theorem 12 can be used as a sufficient condition for approaching the standard problem of AS of DTNUS (1), where the presence of uncertainties (expressed by interval-type coefficients) makes rather difficult the usage of classical procedures.

For practice, it is hard to handle DE (9) in order to check its AS, analytically. A more attractive approach is to just find a sufficient condition for CWAS_v based on an operator with a more tractable form than \mathbf{g} in DE (9).

Theorem 13 *Consider the DE:*

$$\mathbf{z}(t+1) = \widehat{\mathbf{g}}(\mathbf{z}(t)); \widehat{g}_i(\mathbf{z}) = \sum_{j=1}^n \widehat{c}_{ij} z_j^{p_{ij}}, i = \overline{1, n}, \quad (15)$$

where the coefficients \widehat{c}_{ij} are defined by: $\widehat{c}_{ij} = \max \{\bar{c}_{ij}, \tilde{c}_{ij}\}, i, j = \overline{1, n}$.

(i) *If EP $\{0\}$ is AS for DE (15), then EP $\{0\}$ is CWAS_v for DTNUS (1).*

(ii) *In the particular case when the interval-type coefficients a_{ij} of DTNUS (1) satisfy the inequalities given below, for each $i, i = \overline{1, n}$:*

IF p_{ij} -even THEN $(a_{ij}^+ \geq -a_{ij}^- \text{ for all } j \text{ OR } a_{ij}^+ \leq -a_{ij}^- \text{ for all } j)$, (16)

the sufficient condition stated at (i) is also necessary for EP $\{0\}$ of DTNUS (1) to be CWAS_v.

Proof. (i) results from the fact that any PS of DE (15) is also a PS of DI (8). (ii) Whenever (16) is true, one has $\mathbf{g}(\mathbf{z}) = \widehat{\mathbf{g}}(\mathbf{z}), \mathbf{z} \in \mathbb{R}^n$. ■

The advantage of Theorem 13 consists in dealing with only one DE (15), whose coefficients have unique and constant values. Moreover, inequalities (16) may be frequently fulfilled in practical studies, fact which ensures a complete answer to the CWAS_v investigation.

4. Componentwise Exponential Asymptotic Stability

Consider the vector function:

$$\mathbf{v}(t) = \alpha r^t, \alpha = [\alpha_1, \dots, \alpha_n]' \in \mathbb{R}^n, \alpha_i > 0, i = \overline{1, n}, 0 < r < 1. \quad (17)$$

Definition 3 EP {0} of DTNUS (1) is called *componentwise exponentially asymptotically stable* (CWEAS) if there exist a vector $\alpha \in \mathbb{R}^n, \alpha > 0$, and a constant $0 < r < 1$ such that for any $\mathbf{x}(t_0) = \mathbf{x}_0$ with $|\mathbf{x}_0| \leq \alpha r^{t_0}$, the following inequality holds: $|\mathbf{x}(t)| \leq \alpha r^t, t \in [t_0, \infty)$.

Remark 8 Definition 3 can be restated in terms of Definition 2, by taking for $\mathbf{v}(t)$ the particular form given by (17).

Theorem 14 EP {0} of DTNUS (1) is CWEAS iff the following nonlinear algebraic inequalities are compatible (have solutions $\alpha_i > 0, i = \overline{1, n}, 0 < r < 1$):

$$\sum_{j=1}^n \frac{\bar{c}_{ij} \alpha_j^{p_{ij}}}{\alpha_i} \leq r, i = \overline{1, n}, \quad (18-a)$$

$$\sum_{j=1}^n \frac{\tilde{c}_{ij} \alpha_j^{p_{ij}}}{\alpha_i} \leq r, i = \overline{1, n}. \quad (18-b)$$

Proof. *Necessity.* It is immediate if one uses the vector function given by (17) in DIs (7-a) and (7-b) and takes $t = 0$. *Sufficiency.* By induction for $t \in \mathbb{Z}_+$, one can show that functions of type (17) satisfying (18-a) and (18-b) also satisfy DIs (7-a) and (7-b). ■

Theorem 15 EP {0} of DTNUS (1) is CWEAS iff the following nonlinear algebraic inequalities are compatible (have solutions $\alpha_i > 0, i = \overline{1, n}$):

$$\sum_{j=1}^n \frac{\bar{c}_{ij} \alpha_j^{p_{ij}}}{\alpha_i} < 1, i = \overline{1, n}, \quad (19-a)$$

$$\sum_{j=1}^n \frac{\tilde{c}_{ij} \alpha_j^{p_{ij}}}{\alpha_i} < 1, i = \overline{1, n}. \quad (19-b)$$

Proof. This is an immediate consequence of Theorem 14 due to the fact that $0 < r < 1$. ■

Remark 9 Theorems 14 and 15 show that the CWEAS property of EP $\{0\}$ can exist even for exponents $p_{ii} \geq 2$, unlike the corresponding results of the continuous case in ([3]) requesting $p_{ii} = 1$. This difference may be simply explained by comparing the general form of DTNUS (1) and the discretization of the continuous-time systems considered in ([3]), as already suggested by Remark 5.

Conditioning the existence of CWEAS to the values of p_{ii} in DTNUS (1) (as stated in Theorem 10) rises a direct question about the link between CWEAS and CWAS_v.

Theorem 16 *There exists a positive vector $v(t)$ meeting condition (12) such that the EP $\{0\}$ of DTNUS (1) is CWAS_v iff the EP $\{0\}$ is CWEAS.*

Proof. *Sufficiency* is obvious according to Definition 3 and Remark 8. *Necessity* can be proven using Theorem 8 and noticing that for at least one value $t^* \in [t_0, \infty)$ one has $v_i(t^* + 1) \leq v_i(t^*)$, $i = \overline{1, n}$ in DIs (7-a) and (7-b). ■

The nonlinear algebraic inequalities (18-a), (18-b) and (19-a), (19-b) can be written compactly in a *matrix form*, using norm ∞ , by considering the square matrices $\bar{\mathbf{M}}, \tilde{\mathbf{M}} \in \mathbb{R}^{n \times n}$ with the following entries:

$$(\bar{\mathbf{M}})_{ij} = \bar{c}_{ij} \frac{\alpha_j^{p_{ij}}}{\alpha_i}, i, j = \overline{1, n}; \quad (20-a)$$

$$(\tilde{\mathbf{M}})_{ij} = \tilde{c}_{ij} \frac{\alpha_j^{p_{ij}}}{\alpha_i}, i, j = \overline{1, n}. \quad (20-b)$$

Theorem 17 *EP $\{0\}$ of DTNUS (1) is CWEAS iff there exist $\alpha_i > 0$, $i = \overline{1, n}$ and $0 < r < 1$ such that:*

$$\max \left\{ \|\bar{\mathbf{M}}\|_\infty, \|\tilde{\mathbf{M}}\|_\infty \right\} \leq r. \quad (21)$$

Proof. The inequalities (18-a) and (18-b) could be written, using (20-a) and (20-b) in the form (21). ■

Theorem 18 *EP $\{0\}$ of DTNUS (1) is CWEAS iff there exist $\alpha_i > 0$, $i = \overline{1, n}$ such that:*

$$\max \left\{ \|\bar{\mathbf{M}}\|_\infty, \|\tilde{\mathbf{M}}\|_\infty \right\} < 1. \quad (22)$$

Proof. This is a direct consequence of Theorem 17, due to the fact that r is subunitary. ■

Remark 10 Theorems 17 and 18 present the advantage of a more tractable formulation from the computational point of view than inequalities (18-a), (18-b) and (19-a), (19-b), respectively. Thus, the determination of $\alpha_i, i = \overline{1, n}$ and r can be approached as a nonlinear optimization problem with adequate constraints.

As already discussed in the general case of CWAS_v, it might be preferable to use a sufficient condition generated from DE (15) in Theorem 13. Therefore, consider the square matrix $\widehat{\mathbf{P}} \in \mathbb{R}^{n \times n}$, with the following entries:

$$\left(\widehat{\mathbf{P}} \right)_{ij} = \widehat{c}_{ij} \varepsilon^{p_{ij}-1}, \varepsilon > 0, \quad (23)$$

where $\widehat{c}_{ij}, i, j = \overline{1, n}$ are defined in Theorem 13. Denote by $\lambda_{\max}(\widehat{\mathbf{P}})$ the eigenvalue of $\widehat{\mathbf{P}}$ (simple or multiple) with the largest absolute value (spectral radius).

Theorem 19 *If, for a given $\varepsilon > 0$, matrix $\widehat{\mathbf{P}}$ is Schur stable, then the EP {0} of DTNUS (1) is CWEAS for some $0 < \alpha_i \leq \varepsilon, i = \overline{1, n}$, and $\lambda_{\max}(\widehat{\mathbf{P}}) \leq r < 1$.*

Proof. It results from Theorem 14, by replacing all $\alpha_{ij}^{p_{ij}-1} > 0, i, j = \overline{1, n}$ by their common upper bound $\varepsilon^{p_{ij}-1}$, followed by the use of Theorem 17. ■

The advantage of Theorem 19 consists in a quick test on the stability of matrix $\widehat{\mathbf{P}}$ which depends on a single parameter $\varepsilon > 0$.

Remark 11 According to Theorem 13, whenever inequalities 16 are satisfied, the existence of a positive $\varepsilon > 0$ for which matrix $\widehat{\mathbf{P}}$ is Schur stable represents a necessary and sufficient condition for EP {0} of DTNUS (1) to be CWEAS.

5. Conclusions

The exploitation of the FI concepts in analyzing the dynamics of DTNUS (1) gives the possibility to individually monitor each variable of the state vector, yielding a characterization in terms of difference inequalities (Theorems 1, 2, 3). The presence of the interval-type coefficients, although increasing the complexity of the results, responds to

a key problem, frequently encountered in practice, when inherent errors affect the accuracy of the model. The FI results open the way to a componentwise approach to the stability by developing specialized instruments for CWAS analysis (Theorems 8, 9, 12, 13) and CWEAS analysis (Theorems 14, 15, 17, 18, 19), respectively.

Despite a rough similarity between the difference equation of DTNUS (1) and the differential equation describing the class of continuous-time systems considered in ([3]), expecting strictly analogue properties for the solutions in the two cases is meaningless, as reflected by Remarks 1, 2, 5, 9. These remarks are able to outline the importance and the novelty of our study, because the discrete-time dynamics we are dealing with is not the result of uniformly sampling the continuous-time dynamics analyzed in ([3]) (which would lead to a difference equation with unavoidable linear terms, generated by the first order approximation of the derivatives in the differential equation).

References

- [1] Hmamed, A. (1997). "Componentwise stability of 1-D and 2-D linear discrete systems", *Automatica*, 33(9), pp. 1759-1762.
- [2] Pastravanu, O. and Voicu, M. (1999). "Flow invariant rectangular sets and componentwise asymptotic stability of interval matrix systems", Proc. of the 5-th European Control Conference ECC'99, Kalsruhe, CD-ROM.
- [3] Pastravanu, O. and Voicu, M. (2001). "Flow invariance in exploring stability of a class of nonlinear uncertain systems", Proc. of the 6-th European Control Conference ECC'01, Porto, CD-ROM.
- [4] Pavel, H. N. (1984). *Differential Equations: Flow Invariance and Applications. Research Notes in Mathematics*, No. 113, Pitman, Boston.
- [5] Voicu, M. (1984a). "Free Response Characterization via flow-invariance", Prep. of the 9-th World Congress of IFAC, Budapest, Vol. 5, pp. 12-17.
- [6] Voicu, M. (1984b). "Componentwise asymptotic stability of linear constant dynamical systems", *IEEE Trans. Automatic Control*, 29(10), pp. 937-939.
- [7] Voicu, M. (1987). "On the application of the flow-invariance method in the control theory and design", Prep. of th 10-th World Congress of IFAC, Munich, Vol. 8, pp. 364-369.

DIFFERENTIAL PROPERTIES OF LIPSCHITZ, HAMILTONIAN AND CHARACTERISTIC FLOWS

Stefan Mirica

Faculty of Mathematics, University of Bucharest

Academiei 14, 70109 Bucharest, Romania

mirica@math.math.unibuc.ro

Abstract In view of possible applications to Hamilton-Jacobi equations, to optimal control and to differential games, we extend the classical differential properties of smooth Hamiltonian and Characteristic flows to Lipschitzian ones using the contingent derivatives of their components

Keywords: Hamiltonian system and flow, Characteristic system and flow, Lipschitzian mapping, contingent derivative

Mathematics Subject Classification: 34A12; 49J52; 35B37

1. Introduction

The aim of this paper is to generalize in terms of the **contingent derivatives** the *basic differential relation*:

$$DV(t, z).(r, u) = \langle P(t, z), DX(t, z).(r, u) \rangle - r.H(t, X^*(t, z)) - \langle q, u_1 \rangle \quad (1.1)$$

if $z = (\xi, q)$, $u = (u_1, u_2)$, satisfied by the components of a *smooth Characteristic flow* $C^*(., .) := (X^*(., .), V(., .))$, that is uniquely associated to a **smooth Hamiltonian system**:

$$(x', p') = h(t, x, p) := \left(\frac{\partial H}{\partial p}(t, x, p), -\frac{\partial H}{\partial x}(t, x, p) \right), \quad (x(T), p(T)) = z \in D^T \quad (1.2)$$

in the following way: the (smooth) *Hamiltonian flow* $X^*(., .) := (X(., .), P(., .))$: $D_h \subseteq R \times D^T \rightarrow R^n \times R^n$ is defined by the unique, maximal (i.e. non-continuable) solutions $X^*(., z) : I(z) \subseteq R \rightarrow R^n \times R^n$, $z \in D^T := \{(\xi, q); (T, \xi, q) \in D\}$ of the problem (1.2) while the third component is given by the formula:

$$V(t, z) := \int_T^t [\langle P(s, z), \frac{\partial H}{\partial p}(s, X^*(s, z)) \rangle - H(s, X^*(s, z))] ds. \quad (1.3)$$

This type of relations are essential for the construction of classical and generalized "characteristic solutions" of Hamilton-Jacobi equations, for deriving "Hopf-Lax formulas" and are particularly useful in optimal control and differential games.

In this paper the Hamiltonian $H(., ., .) : D = \text{Int}(D) \subseteq R \times R^n \times R^n \rightarrow R$ is differentiable with respect to the last two variables and such that the corresponding *Hamiltonian vector field* $h(., ., .)$ in (1.2) is a "Carathéodory-Lipschitz" mapping; under these hypotheses, if $T \in \text{pr}_1 D$ then the *unique maximal Characteristic flow* $C^*(., .)$ is locally-Lipschitz with respect to the second variable and locally-*AC* (absolutely continuous) with respect to the first one so the basic relation in (1.1) does not make sense any more.

The main result of this paper, Theorem 3.4 below, states that in this case the *contingent derivatives* of the components $X(., .), V(., .)$ satisfy certain relations that coincide with the one in (1.1) in the particular case $h(., ., .)$ is a smooth (Hamiltonian) vector field; for the proof of this result we essentially use the generalization in Blagodatskikh(1973) and Mirică(1985,2002) of the Bendixson-Picard-Lindelöf theorem on *differentiability of solutions with respect to initial data* in the theory of ODE; this type of proof is new and apparently simpler than the traditional proofs of the relation (1.1) in the classical case (e.g. Courant(1962), Hartman(1964), Mirică(1987) etc).

The paper is organized as follows: in Section 2 we present the necessary notations, definitions and preliminary results from Nonsmooth Analysis and from the theory of Carathéodory differential equations and in Section 3 we present the main results.

2. Notations, definitions and preliminary results

From the multitude of the existing concepts in Nonsmooth Analysis (e.g. Aubin and Frankowska(1990), Mirică(1982), etc.), we shall use in the first place the *set-valued contingent directional derivative* of a mapping $f(.) : X \subseteq R^n \rightarrow R^k$ at a point $x \in \text{Int}(X)$ in a direction $u \in R^n$ defined by:

$$K^+ f(x; u) := \{v \in R^k; \exists (s_m, u_m) \rightarrow (0_+, u) : \frac{f(x + s_m u_m) - f(x)}{s_m} \rightarrow v\} \quad (2.1)$$

and, in the case $g(.) : X \subseteq R^n \rightarrow R$ is a real function we may use also its *extreme contingent derivatives (to the right)* at $x \in \text{Int}(X)$ in direction $u \in R^n$:

$$\begin{aligned} \overline{D}_K^+ g(x; u) &:= \limsup_{(s, v) \rightarrow (0_+, u)} \frac{g(x + s.v) - g(x)}{s}, \\ \underline{D}_K^+ g(x; u) &:= \liminf_{(s, v) \rightarrow (0_+, u)} \frac{g(x + s.v) - g(x)}{s} \end{aligned} \quad (2.2)$$

In what follows we shall consider only the particular case in which the domain $X = \text{dom}(f(.)) = \text{dom}(g(.)) \subseteq R^n$ is *open* and the mappings $f(.), g(.)$ are locally-Lipschitz.

We recall first that for this type of mappings the contingent derivatives in (2.1), (2.2) have the properties in the following Proposition whose proof is straightforward (see Aubin and Frankowska(1990), Mirică(1982), etc.):

Proposition 2.1. *If the mapping $f(.) : X = \text{Int}(X) \subseteq R^n \rightarrow R^k$ is locally-Lipschitz and $x \in X$ then its contingent derivative in (2.1) has the following properties:*

(i) *in any direction $u \in R^n$ the subset $K^+ f(x; u) \subset R^k$ is non-empty and compact and is given by:*

$$K^+ f(x; u) := \{v \in R^k; \exists s_m \rightarrow 0_+ : \frac{f(x + s_m u) - f(x)}{s_m} \rightarrow v\}; \quad (2.3)$$

(ii) *the multifunction $K^+ f(x; .)$ is ("globally") Lipschitzian with respect to the Pompeiu-Hausdorff distance in the sense that:*

$$d_H(K^+ f(x; u), K^+ f(x; u')) \leq L \cdot \|u - u'\| \quad \forall u, u' \in R^n \quad (2.4)$$

where L is the Lipschitz constant of $f(\cdot)$ at x and:

$$d_H(A, B) := \max\{d^*(A, B), d^*(B, A)\}, \quad d^*(A, B) := \sup_{a \in A} \inf_{b \in B} \|a - b\|.$$

Moreover, $K^+ f(x; \cdot)$ is positively homogeneous in the sense that:

$$K^+ f(x; \lambda \cdot u) = \lambda K^+ f(x; u) \quad \forall u \in R^n, \lambda \geq 0, \quad K^+ f(x; 0) = \{0\};$$

(iii) the mapping $f(\cdot)$ is (Fréchet) differentiable at the point $x \in X$ iff it is contingent differentiable in any direction $u \in R^n$ in the sense that:

$$\exists f_K^+(x; u) := \lim_{(s, v) \rightarrow (0_+, u)} \frac{f(x + s \cdot v) - f(x)}{s} \quad \forall u \in R^n \quad (2.5)$$

and the "contingent derivative" $f_K^+(x; \cdot) : R^n \rightarrow R^k$ is linear; in this case the two derivatives coincide i.e. $Df(x) = f_K^+(x; \cdot)$.

(iv) if $g(\cdot) : X \rightarrow R$ is locally-Lipschitz then its contingent derivatives in (2.1)-(2.2) are related as follows:

$$\bar{D}_K^+ g(x; u) = \max[K^+ g(x; u)] \geq \min[K^+ g(x; u)] = \underline{D}_K^+ g(x; u) \quad \forall u \in R^n. \quad (2.6)$$

Remark 2.2. According to the well-known Rademacher's theorem, the locally-Lipschitz mapping $f(\cdot)$ in Prop.2.1 is a.e. differentiable hence there exists a subset $\mathcal{D}_F(f) \subseteq X$, of "full Lebesgue measure" ($\mu(\cdot)$) such that:

$$\exists Df(x) = f_K^+(x; \cdot) \quad \forall x \in \mathcal{D}_F(f), \quad \mu(X \setminus \mathcal{D}(f)) = 0; \quad (2.7)$$

on the other hand, as simple examples show, the set of points at which the contingent derivative in (2.6) exists may be strictly larger than the (Fréchet) "differentiability set" in (2.7):

$$\mathcal{D}_F(f) \subseteq \bar{D}_K^+(f) := \{x \in X; \text{dom}(f_K^+(x; \cdot)) = R^n\}; \quad (2.8)$$

equivalently, for each point $x \in X$ one may consider the "set of contingent differentiability directions" of $f(\cdot)$ at x defined as the domain of the mapping $f_K^+(x; \cdot)$:

$$\mathcal{D}_K^+(f; x) := \text{dom}(f_K^+(x; \cdot)) := \{u \in R^n; \exists f_K^+(x; u)\} \quad (2.9)$$

so that one may write: $\mathcal{D}_K^+(f) = \{x \in X; \mathcal{D}_K^+(f; x) = R^n\}$.

In this paper we shall use these concepts and results to the study of the "maximal" (i.e. non-continuable) flow, $x_f(\cdot, \cdot, \cdot) : D_f \subseteq R \times D \rightarrow R^n$, of a Carathéodory-Lipschitz differential equation:

$$x' = f(t, x), \quad x(s) = y, \quad (s, y) \in D = \text{dom}(f(\cdot, \cdot)) \subseteq R \times R^n \quad (2.10)$$

defined, more precisely, as follows:

Definition 2.3. A mapping $f(\cdot, \cdot) : D \rightarrow R^n$ is said to be a *Carathéodory-Lipschitz (C-L) vector field* if $D \subset R \times R^n$ is open and the following properties hold:

(i) $f(\cdot, \cdot)$ is a *Carathéodory mapping* in the sense that the mappings $f(\cdot, x)$, $x \in pr_2 D$ are (Lebesgue) measurable, there exists a null subset $I_f \subset pr_1 D$ (i.e. $\mu(I_f) = 0$) such that $f(t, \cdot)$, $t \in pr_1 D \setminus I_f$ are continuous and, moreover, $f(\cdot, \cdot)$ is *locally integrably bounded* in the sense that for any compact subset $D_0 \subset D$ there exist an integrable function $m(\cdot) \in L^1(pr_1 D_0; R_+)$, $R_+ = [0, \infty)$ and a null subset $I_0 \subset pr_1 D_0$ such that:

$$\|f(t, x)\| \leq m(t) \quad \forall (t, x) \in D_0, \quad t \in pr_1 D_0 \setminus I_0; \quad (2.11)$$

(ii) $f(.,.)$ is *locally-integrably Lipschitz with respect to the second variable* in the sense that for any compact subset $D_0 \subset D$ there exist an integrable function $L(.) \in L^1(pr_1 D_0; R_+)$ and a null subset $I_0 \subset pr_1 D_0$ such that:

$$\|f(t, x) - f(t, y)\| \leq L(t) \|x - y\| \quad \forall (t, x), (t, y) \in D_0, t \in pr_1 D_0 \setminus I_0. \quad (2.12)$$

Remark 2.4. As particular cases of the **C-L** vector fields in Def.2.3 one may consider the *locally-essentially bounded* ones when the function $m(.)$ in (2.11) is essentially bounded (i.e. $m(.) \in L^\infty(pr_1 D_0; R_+)$), the *essentially bounded locally-Lipchitz* ones for which $L(.) \in L^\infty(pr_1 D_0; R_+)$ and the *Carathéodory-C¹* vector fields for which, in addition to the properties in Def.2.3, one assumes that the mappings $f(t, .)$, $t \in pr_1 D \setminus I_f$ are differentiable and, moreover, the derivative, $D_2 f(.,.)$ is of Carathéodory type i.e. $D_2 f(., x)$, $x \in pr_2 D$ are measurable, the mappings $D_2 f(t, .)$, $t \in pr_1 D \setminus I_f$ are continuous and for any compact subset $D_0 \subset D$ there exist $L(.) \in L^1(pr_1 D_0; R_+)$ and a null subset $I_0 \subset pr_1 D_0$ such that:

$$\|D_2 f(t, x)\| \leq L(t) \quad \forall (t, x), (t, y) \in D_0, t \in pr_1 D_0 \setminus I_0. \quad (2.13)$$

As further particular cases one may consider *Peano-Lipschitz vector fields* $f(.,.)$ which are continuous with respect to both variables and locally-Lipschitz with respect to the second variable, uniformly with respect to the first one (i.e. the function $L(.)$ in (2.12) is constant) and the "classical" *Peano-C¹* vector fields for which both mappings $f(., .)$, $D_2 f(., .)$ are continuous.

We summarize the basic results of the theory of Carathéodory Ordinary Differential Equations in the following theorem for whose proof we refer to Kurzweil(1986),Ch.18:

Theorem 2.5. *If $f(., .)$ is a Carathéodory-Lipschitz (**C-L**) vector field in the sense of Def.2.3 then the following statements hold:*

(i) *for any initial point $(s, y) \in D$ there exists a unique maximal (i.e. non-continuable) locally-AC (absolutely continuous on each compact interval) Carathéodory solution $x_f(.; s, y) : I(s, y) \subset R \rightarrow R^n$ of the equation in (2.10) that satisfies:*

$$D_1 x_f(t; s, y) = f(t, x_f(t; s, y)) \text{ a.e.} (t \in I(s, y)), \quad x_f(s; s, y) = y; \quad (2.14)$$

(ii) *moreover, $I(s, y)$ is an open interval containing s , the domain of the flow, $D_f = \{(t, s, y); (s, y) \in D, t \in I(s, y)\} \subseteq R \times D$ is an open subset, and the "maximal flow" $x_f(.; .) : D_f \rightarrow R^n$ is continuous;*

(iii) *the mappings $x_f(t; s, .)$ are locally Lipschitz and the mappings $x_f(.; s, y)$, $x_f(t; ., y)$ are locally-AC and satisfy the equivalent "associated integral equation"*

$$x_f(t; s, y) = y + \int_s^t f(\sigma, x_f(\sigma; s, y)) d\sigma \quad \forall (t, s, y) \in D_f. \quad (2.15)$$

Therefore, each of the mappings, $x_f(.; s, y)$, $(s, y) \in D$ satisfies the equation (2.10) outside of a null subset of the interval $I(s, y)$; however, the following important result shows that equation (2.10) is satisfied outside a "common" null subset by all the solutions:

Theorem 2.6 (Scorza-Dragoni(1948)). *If $f(., .)$ is a **C-L** vector field in the sense of Def.2.3 then there exists a null subset $J_f \subset pr_1 D$ such that:*

$$D_1 x_f(t; s, y) = f(t, x_f(t; s, y)) \quad \forall t \in I(s, y) \setminus J_f, (s, y) \in D. \quad (2.16)$$

For a proof of this theorem (actually valid for the more general class of the Carathéodory vector fields in Def.2.3) one may see also Th.18.4.9 in Kurzweil (1986).

The differentiability properties of the flow $x_f(\cdot; \cdot, \cdot)$ in Th.2.5 will be expressed, as usual, along a (fixed) "reference trajectory", $z(\cdot)$, given by:

$$z(t) = x_f(t; t_0, x_0), \quad t \in I(t_0, x_0) \quad (2.17)$$

where $(t_0, x_0) \in D$ is a fixed point, in terms of the set-valued contingent directional derivatives in (2.1) of the mappings $f(t, \cdot)$, $x_f(t; t_0, \cdot)$, $x_f(t, \cdot, x_0)$ and of the Carathéodory solutions, $v(\cdot)$, of the **contingent variational inclusion (CVI)**:

$$v'(t) \in co[K^+ f(t; \cdot)(z(t); v(t))] \text{ a.e.}(I(t_0, x_0)), \quad v(t_0) = u_0 \in R^n \quad (2.18)$$

where $co[A]$ denotes the *convex hull of the subset* $A \subset R^n$ which, as it is well-known, coincides with the *closed convex hull*, $\bar{co}[A]$, whenever A is compact; using the "contingent differentiability directions" in (2.9) one may see that at certain points the **CVI** in (2.18) becomes the **contingent variational equation (CVE)**:

$$v'(t) = (f(t; \cdot))_K^+(z(t); v(t)) \text{ a.e.}(I(t_0, x_0)), \quad \text{if } v(t) \in \mathcal{D}_K^+(f(t; \cdot); z(t)), \quad (2.19)$$

while on the ("full measure") Fréchet differentiability sets in (2.7) the **CVI** in (2.18) becomes the classical **variational equation (VE)**:

$$v'(t) = D_2 f(t, z(t)).v(t) \text{ a.e.}(I(t_0, x_0)), \quad \text{if } z(t) \in \mathcal{D}_F(f(t, \cdot)). \quad (2.20)$$

In the proof of the main result in the next section we shall essentially uses the following generalization of the classical Bendixson-Picard- Lindelöf theorem:

Theorem 2.7. Let $f(\cdot, \cdot) : D \rightarrow R^n$ be a Carathéodory- Lipschitz vector field in the sense of Def.2.3, let $(t_0, x_0) \in D$, $t_1 \in I(t_0, x_0)$, $I_1 := [t_0, t_1] \subset I(t_0, x_0)$ and let $z(\cdot)$ be the reference trajectory in (2.17).

Then the contingent derivatives in (2.1) of the maximal flow $x_f(\cdot, \cdot, \cdot)$ in Th.2.5 have the following properties:

(i) for any vectors $u_0 \in R^n$, $u_1 \in K^+ x_f(t_1; t_0, \cdot)(x_0; u_0)$ there exists a Carathéodory solution $v(\cdot) : I_1 \rightarrow R^n$ of the contingent variational inclusion **CVI** such that:

$$v(t_0) = u_0, \quad v(t_1) = u_1, \quad v(t) \in K^+ x_f(t; t_0, \cdot)(x_0; u_0) \quad \forall t \in I_1 \quad (2.21)$$

(ii) if $J_f \subset pr_1 D$ is the null subset in (2.16) and $t_0 \in pr_1 D \setminus J_f$ then for any vector $u_1^0 \in K^+ x_f(t_1; \cdot, x_0)(t_0; 1)$ there exists a Carathéodory solution $v^0(\cdot)$ of the contingent variational inclusion **CVI** in (2.18) such that:

$$v^0(t_0) = -f(t_0, x_0), \quad v^0(t_1) = u_1^0, \quad v^0(t) \in K^+ x_f(t; \cdot, x_0)(t_0; 1) \quad \forall t \in I_1. \quad (2.22)$$

Moreover, if $f(\cdot, \cdot)$ is locally essentially bounded in the sense of Remark 2.4 and $t_0 \in J_f$ then for any vector $u_1^0 \in K^+ x_f(t_1; \cdot, x_0)(t_0; 1)$ there exists a Carathéodory solution $v^0(\cdot)$ of the contingent variational inclusion **CVI** in (2.18) that satisfies the last two conditions in (2.22) and also the "weaker" initial condition:

$$v^0(t_0) \in -f^{co}(t_0+, x_0), \quad f^{co}(s+, y) := \bigcap_{\delta > 0} \bigcap_{\mu(J)=0} \bar{co}f([s, s + \delta] \setminus J) \times B_\delta(y). \quad (2.23)$$

For a proof of this result we refer to Blagodatskikh(1973) and Mirică(1985,2002).

3. The main results

Everywhere in what follows we assume the following:

Hypothesis 3.1. The Hamiltonian $H(\cdot, \cdot, \cdot) : D = \text{Int}(D) \subseteq R \times R^n \times R^n \rightarrow R$ is such that there exists a null subset, $I_H \subset \text{pr}_1 D$ (i.e. $\mu(I_H) = 0$) such that the functions $H(t, \cdot, \cdot), t \in \text{pr}_1 D \setminus I_H$ are (Fréchet) differentiable and such that the corresponding *Hamiltonian vector field* $h(\cdot, \cdot, \cdot) = (h_1(\cdot, \cdot, \cdot), h_2(\cdot, \cdot, \cdot))$ in (1.2) is a *Carathéodory-Lipschitz vector field* in the sense of Def.2.3; moreover, the null subset $I_H \subset \text{pr}_1 D$ is taken such that for each $t \in \text{pr}_1 D \setminus I_H$ the mapping $h(t, \cdot)$ is locally-Lipschitz (on the "section" $D^t := \{z \in R^{2n}; (t, z) \in D\}$).

To simplify the exposition we consider here only the particular case in which the initial values of the "time-variable" in (1.2) are fixed though statement (ii) in Th.2.7 suggests the possibility of studying also the more general case in which these values are variable; noting first that the **vector field of the characteristics** defined by:

$$\begin{aligned} c(t, x, p) &:= (h_1(t, x, p), h_2(t, x, p), c_3(t, x, p)), \quad h_1(t, x, p) := \frac{\partial H}{\partial p}(t, x, p), \\ h_2(t, x, p) &:= -\frac{\partial H}{\partial x}(t, x, p), \quad c_3(t, x, p) := \langle p, h_1(t, x, p) \rangle - H(t, x, p) \end{aligned} \quad (3.1)$$

is also of Carathéodory-Lipschitz type, applying Ths.2.5, 2.6 one obtains:

Theorem 3.2. If Hypothesis 3.1 is satisfied, $T \in \text{pr}_1 D$ and $D^T := \{(\xi, q) \in R^n \times R^n; (T, \xi, q) \in D\}$ then there exists a unique characteristic flow $C^*(\cdot, \cdot) = (X^*(\cdot, \cdot), V(\cdot, \cdot)) : D_h \subseteq R \times D^T \rightarrow R^n \times R^n \times R$ and a null subset $J_c \subset \text{pr}_1 D$, ($I_H \subseteq J_c$) such that for each $z = (\xi, q) \in D^T$ the mapping $C^*(\cdot, z) : I(z) \subseteq R \rightarrow R^n \times R^n \times R$ is the unique maximal Carathéodory solution of the "system of characteristics":

$$(x', p', v') = c(t, x, p), \quad (x(T), p(T), v(T)) = (\xi, q, 0) \in D^T \times \{0\}$$

satisfying:

$$D_1 C^*(t, z) = c(t, X^*(t, z)) \quad \forall t \in I(z) \setminus J_c, \quad C^*(T, z) = (z, 0), \quad z \in D^T. \quad (3.2)$$

Moreover, the characteristic flow $C^*(\cdot, \cdot)$ has the regularity properties in Th.2.5 i.e. the intervals $I(z) \subseteq R$ and the domain $D_h := \{(t, z); z \in D^T, t \in I(z)\}$ are open, $C^*(\cdot, \cdot)$ is continuous, the mappings $C^*(t, \cdot)$ are Lipschitzian and the mappings $C^*(\cdot, z)$ are locally-AC.

In addition, $X^*(\cdot, \cdot) = (X(\cdot, \cdot), P(\cdot, \cdot))$ is the unique maximal flow in the same sense of the Hamiltonian system in (1.2) and the third component, $V(\cdot, \cdot)$, is given by the formula in (1.3).

Before applying Th.2.7 to the characteristic flow above we prove first a very specific property of the contingent derivatives in (2.1) of the characteristic vector field in (3.1).

Proposition 3.3. If $I_H \subset \text{pr}_1 D$ is the null subset in Hypothesis 3.1 then at any point $(t, z) \in D$, $t \in \text{pr}_1 D \setminus I_H$, $z = (x, p) \in D^t$ and in any direction $u = (u_1, u_2) \in R^{2n}$, the set-valued contingent derivative in (2.1) of the characteristic vector field in (3.1) is given by:

$$\begin{aligned} K^+ c(t, \cdot)(z; u) &= \{v = (v_1, v_2, v_3); (v_1, v_2) \in K^+ h(t, \cdot)(z; u), \\ v_3 &= \langle p, v_1 \rangle + \langle h_2(t, z), u_1 \rangle\} \text{ if } z = (x, p) \in D^t, \quad t \in \text{pr}_1 D \setminus I_H \end{aligned} \quad (3.3)$$

and its (closed) convex hull is given by:

$$\begin{aligned} \text{co}[K^+ c(t, \cdot)(z; u)] &= \{v = (v_1, v_2, v_3); (v_1, v_2) \in \text{co}[K^+ h(t, \cdot)(z; u)], \\ v_3 &= \langle p, v_1 \rangle + \langle h_2(t, z), u_1 \rangle\}. \end{aligned} \quad (3.4)$$

where $h = (h_1, h_2)$ is the Hamiltonian vector field in (3.1).

In particular, at the Fréchet differentiability points $z = (x, p) \in \mathcal{D}_F(h(t, .))$ in (2.7), the component $c_3(t, .)$ is also differentiable and:

$$D_2 c_3(t, z).u = \langle p, D_2 h_1(t, z).u \rangle + \langle h_2(t, z), u_1 \rangle \quad \forall u = (u_1, u_2) \in R^n \times R^n. \quad (3.5)$$

Proof. From the equivalent definition in (2.3) (for locally-Lipschitz mappings) it follows that if $v = (v_1, v_2, v_3) \in K^+ c(t, .)(z; u)$ then there exists a sequence $s_m \rightarrow 0_+$ such that:

$$(v_1, v_2) = \lim_{m \rightarrow \infty} \frac{h(t, z + s_m u) - h(t, z)}{s_m}, \quad v_3 = \lim_{m \rightarrow \infty} \frac{c_3(t, z + s_m u) - c_3(t, z)}{s_m} \quad (3.6)$$

hence $(v_1, v_2) \in K^+ h(t, .)(z; u)$; on the other hand, from (3.6), (3.1) it follows:

$$v_3 = \lim_{m \rightarrow \infty} \left[\langle p, \frac{h_1(t, z + s_m u) - h_1(t, z)}{s_m} \rangle + \langle u_2, h_1(t, z + s_m u) \rangle - \frac{H(t, z + s_m u) - H(t, z)}{s_m} \right]$$

and since $H(t, .)$ is differentiable, from (3.1) it follows:

$$\lim_{m \rightarrow \infty} \frac{H(t, z + s_m u) - H(t, z)}{s_m} = -\langle h_2(t, z), u_1 \rangle + \langle h_1(t, z), u_2 \rangle$$

hence $v_3 = \langle p, v_1 \rangle + \langle u_2, h_1(t, z) \rangle + \langle h_2(t, z), u_1 \rangle - \langle h_1(t, z), u_2 \rangle$ and the first inclusion ("⊆") in (3.3) is proved.

To prove the reversed inclusion we consider $v = (v_1, v_2, v_3)$ such that $(v_1, v_2) \in K^+ h(t, .)(z; u)$, $v_3 = \langle p, v_1 \rangle + \langle h_2(t, z), u_1 \rangle$ and note that from (2.3) it follows the existence of a sequence $s_m \rightarrow 0_+$ such that the first relation in (3.6) is verified; next, since the component $c_3(t, .)$ in (3.1) is obviously locally-Lipschitz, the sequence

$$\left\{ \frac{c_3(t, z + s_m u) - c_3(t, z)}{s_m}, m \in N \right\} \subset R$$

is bounded hence it has a convergent subsequence and therefore there exists $\bar{v}_3 \in R$ such that, taking possibly a subsequence, one has:

$$(v_1, v_2, \bar{v}_3) = \lim_{m \rightarrow \infty} \frac{c(t, z + s_m u) - c(t, z)}{s_m} \in K^+ c(t, .)(z; u).$$

Finally, from the proof above it follows that in this case $\bar{v}_3 = \langle p, v_1 \rangle + \langle h_2(t, z), u_1 \rangle = v_3$ and the relation in (3.3) is proved.

The relation in (3.4) follows, obviously from the one in (3.3) since v_3 depends "linearly" on v_1 while, in view of Prop.2.1, (3.5) is a particular case of (3.3),(3.4).

We note that the relation in (3.4), which seems to be ignored in the classical theory, may be extended in the same form to the "sets of contingent differentiable directions" $u \in \mathcal{D}_K^+(h(t, .); z)$ defined in (2.9).

The main result of this paper is the following:

Theorem 3.4. *If Hypothesis 3.1 is satisfied and $C^*(., .) = (X(., .), P(., .), V(., .))$ is the characteristic flow in Th.3.2 then at each point $(t, z) \in D_h$, $z = (\xi, q) \in D^T$ and any direction $u = (u_1, u_2) \in R^n \times R^n$, the contingent derivatives in (2.1) with respect to the second variable of its components are related as follows:*

$$K^+ V(t, .)(z; u) = \{ \langle P(t, z), v_1 \rangle - \langle q, u_1 \rangle; v_1 \in K^+ X(t, .)(z; u) \} \quad (3.7)$$

and therefore the extreme contingent derivatives in (2.2), (2.6) are given by:

$$\begin{aligned}\bar{D}_K^+ V(t, \cdot)(z; u) &= \max\{\langle P(t, z), v_1 \rangle - \langle q, u_1 \rangle; v_1 \in K^+ X(t, \cdot)(z; u)\} \\ \underline{D}_K^+ V(t, \cdot)(z; u) &= \min\{\langle P(t, z), v_1 \rangle - \langle q, u_1 \rangle; v_1 \in K^+ X(t, \cdot)(z; u)\}.\end{aligned}\quad (3.8)$$

In particular, if $u \in \mathcal{D}_K^+(X(t, \cdot); z)$ (i.e. $X(t, \cdot)$ is contingent differentiable at z in direction u) then $u \in \mathcal{D}_K^+(V(t, \cdot); z)$ and:

$$(V(t, \cdot))_K^+(z; u) = \langle P(t, z), (X(t, \cdot))_K^+(z; u) \rangle - \langle q, u_1 \rangle \text{ if } u = (u_1, u_2) \quad (3.9)$$

and if $z \in \mathcal{D}_F(X(t, \cdot))$ (i.e. $X(t, \cdot)$ is differentiable at z) then $V(t, \cdot)$ is also differentiable at z and:

$$D_2 V(t, z).u = \langle P(t, z), D_2 X(t, z).u \rangle - \langle q, u_1 \rangle \quad \forall u = (u_1, u_2) \in R^{2n}. \quad (3.10)$$

Proof. To prove the inclusion " \subseteq " in (3.7) we consider $v_3 \in K^+ V(t, \cdot)(z; u)$ and note that from (2.3) it follows that there exists $s_m \rightarrow 0_+$ such that:

$$v_3 = \lim_{m \rightarrow \infty} \frac{V(t, z + s_m u) - V(t, z)}{s_m}; \quad (3.11)$$

next, since $C^*(t, \cdot) = (X(t, \cdot), P(t, \cdot), V(t, \cdot))$ is locally-Lipschitz, as in the case above it follows that there exists $(v_1, v_2) \in K^+ X^*(t, \cdot)(z; u)$ such that, taking possibly a subsequence, the relations in (3.6) hold.

In order to prove that in this case one has $v_3 = \langle P(t, z), v_1 \rangle - \langle q, u_1 \rangle$ we apply Th.2.7 to the **C-L** "standard" vector field defined by:

$$\tilde{c}(t, x, p, v) := c(t, x, p) \quad \forall (t, x, p) \in D, v \in R \quad (3.12)$$

and to its corresponding maximal flow, $\tilde{C}^*(\cdot, \cdot)$ for which one obviously has:

$$\tilde{C}^*(t, \tilde{z}) = C^*(t, z) \text{ if } \tilde{z} = (z, 0) \in D^T \times \{0\}. \quad (3.13)$$

We take the "reference trajectory"

$$C(s) := (X(s), P(s), V(s)) := \tilde{C}^*(s, \tilde{z}) = C^*(s, z), \quad s \in I_1 = [T, t] \quad (3.14)$$

and note that from statement (i) of Th.2.7 it follows that for the vectors $\tilde{u} := (u, 0) \in R^{2n} \times \{0\}$, $v \in K^+ \tilde{C}^*(t, \cdot)(\tilde{z}; \tilde{u})$ there exists a Carathéodory solution $w(\cdot) = (w_1(\cdot), w_2(\cdot), w_3(\cdot)) \in AC(I; R^{2n+1})$ of the **CVI**:

$$w'(s) \in co[K^+ \tilde{c}(s, \cdot)(C(s); w(s))] \text{ a.e.}(I_1) \quad (3.15)$$

such that:

$$w(T) = \tilde{u}, \quad w(t) = v, \quad w(s) \in K^+ \tilde{C}^*(s, \cdot)(\tilde{z}; \tilde{u}) \quad \forall s \in I_1 = [T, t]. \quad (3.16)$$

We note that from (3.12) and (2.1) it follows that:

$$K^+ \tilde{c}(t, \cdot)(\tilde{z}, \tilde{u}) = K^+ c(t, \cdot)(z; u) \text{ if } \tilde{z} = (z, r) \in D^t \times R, \quad \tilde{u} = (u, 0) \in R^{2n} \times \{0\}$$

hence the **CVI** in (3.15) becomes:

$$w'(s) \in co[K^+ c(s, \cdot)(X^*(s); (w_1(s), w_2(s)))] \text{ a.e.}(I_1), \quad X^*(\cdot) = (X(\cdot), P(\cdot))$$

We apply now Prop.3.3 to conclude that one has:

$$\begin{aligned} (w'_1(s), w'_2(s)) &\in \text{co}[K^+ h(s, \cdot)(X^*(s); w_1(s), w_2(s))] \text{ a.e.}(I_1) \\ w'_3(s) &= \langle P(s), w'_1(s) \rangle + \langle h_2(s, X^*(s)), w_1(s) \rangle \text{ a.e.}(I_1). \end{aligned} \quad (3.17)$$

Therefore, since from (1.2) it follows that $P'(s) = h_2(s, X^*(s))$ a.e.(I_1), from the last relation in (3.17) it follows that:

$$w'_3(s) = \frac{d}{ds} [\langle P(s), w_1(s) \rangle] \text{ a.e.}(I_1)$$

hence using the Leibnitz-Newton formula for AC mappings one obtains: $w_3(t) = w_3(T) + \langle P(t), w_1(t) \rangle - \langle P(T), w_1(T) \rangle$ which, together with the end-point conditions in (3.16) and the fact that $X^*(T) = (\xi, q) = z$, proves the fact that $v_3 = \langle P(t, z), v_1 \rangle - \langle q, u_1 \rangle$ and the first inclusion in (3.7) is proved.

To prove the reversed inclusion we consider $v_1 \in K^+ X(t, \cdot)(z; u)$, $v_3 := \langle P(t, z), v_1 \rangle - \langle q, u_1 \rangle$ and note that from (2.3) it follows that there exists $s_m \rightarrow 0_+$ such that: $v_1 = \lim_{m \rightarrow \infty} (X(t, z + s_m u) - X(t, z))/s_m$; as in the other cases above, since $V(t, \cdot)$ is locally-Lipschitz it follows that there exists $\bar{v}_3 \in R$ such that, taking possibly a subsequence one has:

$$\bar{v}_3 = \lim_{m \rightarrow \infty} \frac{V(t, z + s_m u) - V(t, z)}{s_m} \in K^+ V(t, \cdot)(z; u)$$

hence from the proof above it follows that $\bar{v}_3 = \langle P(t, z), v_1 \rangle - \langle q, u_1 \rangle = v_3$ and the theorem is completely proved since the relations in (3.9) and (3.10) are obvious particular cases of the one in (3.7).

Noting that for $r = 0$ the relation in (1.1) coincide with (3.10), we prove now a more complete generalization of (1.1) in the case $r \in R$:

Corollary 3.5. *If Hypothesis 3.1 is satisfied and $J_c \subset \text{pr}_1 D$ is the null subset in (3.2) then at each point $z = (\xi, q) \in D^T$, $t \in I(z) \setminus J_c$ and any direction $(r, u) \in R \times R^{2n}$, the contingent derivatives in (2.1) of the components of the characteristic flow are related as follows:*

$$\begin{aligned} K^+ V((t, z); (r, u)) &= \{\langle P(t, z), v_1 \rangle - r.H(t, X^*(t, z)) - \langle q, u_1 \rangle; \\ v_1 &\in K^+ X((t, z); (r, u))\}, \text{ if } u = (u_1, u_2), t \in I(z) \setminus J_c, z \in D^T. \end{aligned} \quad (3.18)$$

In particular, if $X(\cdot, \cdot)$ is differentiable at $(t, z) \in D_h$ then $V(\cdot, \cdot)$ is also differentiable at the same point and the formula in (1.1) is verified.

Proof. We prove first that outside of the null subset J_c in (3.2) the contingent derivatives of the characteristic flow $C^*(\cdot, \cdot)$ (and therefore, of each of its components) satisfy the relation:

$$K^+ C^*((t, z); (r, u)) = D_1 C^*(t, z).r + K^+ C^*(t, \cdot)(z; u) \text{ if } t \in I(z) \setminus J_c. \quad (3.19)$$

To prove the inclusion " \subseteq " we consider $v \in K^+ C^*((t, z); (r, u))$ and note that from (2.1) it follows that there exists a sequence $(s_m, r_m, u_m) \rightarrow (0_+, r, u)$ such that:

$$v = \lim_{m \rightarrow \infty} \frac{1}{s_m} [C^*(t + s_m r_m, z + s_m u_m) - C^*(t, z)]; \quad (3.20)$$

next, since $C^*(t, \cdot)$ is Lipschitzean, taking possibly a subsequence one may assume that:

$$\begin{aligned} \exists v^2 &:= \lim_{m \rightarrow \infty} \frac{1}{s_m} [C^*(t, z + s_m u_m) - C^*(t, z)] \in K^+ C^*(t, \cdot)(z; u) \\ v - v^2 &= v^1 := \lim_{m \rightarrow \infty} \frac{1}{s_m} [C^*(t + s_m r_m, z + s_m u_m) - C(t, z + s_m u_m)]. \end{aligned} \quad (3.21)$$

Using the "integrable-Lipschitz" property in Def.2.3 of $c(., .)$ and the "usual" Lipschitz property in Th.3.2 of $C^*(., .)$ one may easily prove that:

$$\begin{aligned} v^1 &:= \lim_{m \rightarrow \infty} \frac{1}{s_m} \int_t^{t+s_m r_m} c(s, X^*(s, z + s_m u_m)) ds = \\ &= \lim_{m \rightarrow \infty} \frac{1}{s_m} \int_t^{t+s_m r_m} c(s, X^*(s, z)) ds = D_1 C^*(t, z).r \end{aligned} \quad (3.22)$$

(since $t \in I(z) \setminus J_c$) and the first inclusion in (3.19) is proved; the reversed inclusion follows in the same way noting that if $v^2 \in K^+ C^*(t, .)(z; u)$ is of the form in (3.21) and $v^1 := D_1 C^*(t, z).r$ then using (3.22) it follows that $v := v^1 + v^2$ is of the form in (3.20) and (3.19) is proved; the relation in (3.18) is an obvious consequence of (3.7) and (3.19) and Cor.3.5 is completely proved.

We note that, apparently, the most general case in which (3.18) is verified also at the points $t \in J_c$ seems to be that in which the mappings $H(., z)$, $h(., z)$, $z \in pr_2 D$ are *regular*, having one-sided limits at each point hence an at most countable number of discontinuity points, all of the first kind.

References

- [1] Aubin, J. P. and H. Frankowska. (1990). *Set Valued Analysis*, Boston, Birkhäuser.
- [2] Blagodatskikh, V. I. (1973). *On differentiability of solutions with respect to initial conditions*, Diff. Uravn., **9**, 2136-2140 (in Russian).
- [3] Courant, R. (1962). *Partial Differential Equations*, N.Y., Interscience.
- [4] Hartman, Ph. (1964) *Ordinary Differential Equations*, N.Y., Wiley
- [5] Kurzweil, J. (1986) *Ordinary Differential Equations*, Amsterdam, Elsevier,.
- [6] Mirică, Șt. (1982). *The contingent and the paratingent as generalized derivatives of vector-valued and set-valued mappings*, Nonlinear Anal., Theory, Meth. Appl., **6**, 1335-1368.
- [7] Mirică, Șt. (1985). *On some generalizations of the Bendixson-Picard theorem in the theory of differential equations*, Bull. Math. Soc. Sci. Math. Roumanie, **29(77)**, 315-328.
- [8] Mirică, Șt. (1987). *Generalized solutions by Cauchy's Method of Characteristics*, Rend. Sem. Mat. Univ. Padova, **77**, 317-350.
- [9] Mirică, Șt. (1995). *Quasitangent differentiability with respect to initial data for Carathéodory-Lipschitz differential equations*, in "Qualitative problems for Differential Equations and Control Theory", C. Corduneanu Ed., New Jersey, Singapore, World Scientific, 81-89.
- [10] Mirică, Șt. (2002). *On differentiability with respect to initial data in the theory of Differential Equations*, Revue Roum. Math. Pure Appl., submitted.
- [11] Mirică, Șt. and C. Neculaescu (1998). *On a semi-smooth Hamiltonian system and related Hamilton-Jacobi equations*, Anal. Univ. "Ovidius", Constanța, Ser. Mat. **6**.
- [12] Scorza-Dragoni, G. (1948). *Una teorema sulla funzione continue rispetto ad una e misurabile rispetto ad un'altra variabile*, Rend. Sem. Mat. Univ. Padova, **XVII**, 102-106.

THE CONTROL OF SAFFMAN-TAYLOR INSTABILITY

Gelu Paşa

Institute of Mathematics of the Romanian Academy

P.O. Box 1-764, RO-70700, Bucharest, Romania

gpasa@stoilow.imar.ro

1. The Saffman–Taylor formula

The Secondary Oil Recovery process is considered: the oil contained in a two dimensional homogeneous porous medium is obtained by injection of an immiscible second (less viscous) fluid. The Hele-Shaw approximation is considered. The instability of the sharp interface between the fluids appears, first studied by Saffman and Taylor, [3]. The (constant) water and oil viscosities are denoted by μ_1 and μ_2 . The velocity components are (u, v) . The perturbations of the horizontal velocity are

$$u'(x, y, t) = f(x) \cdot \exp(iky + \sigma t),$$

where σ is the growth constant and k is the wave number in the Oy direction. The Saffman - Taylor value for the growth constant is

$$\sigma = \frac{(\mu_2 - \mu_1)Uk - Tk^3}{\mu_2 + \mu_1},$$

where T is the surface tension on $x = 0$ (the water - oil interface). A maximal value σ_m (in terms of k) is obtained for the wave number k_m :

$$\sigma_m = \sigma_m(k_m) = \frac{2(\mu_2 - \mu_1)U}{3(\mu_2 + \mu_1)\sqrt{3}} \cdot \sqrt{\frac{(\mu_2 - \mu_1)U}{T}}, \quad k_m = \frac{1}{\sqrt{3}} \sqrt{\frac{(\mu_2 - \mu_1)U}{T}}.$$

We use in the sequel the dimensionless quantities given in [3]:

$$\begin{cases} \sigma^* = \frac{2\sigma}{3\sqrt{3}\sigma_m}, & \mu^*(x) = \mu_0(x)/\mu_1, \quad k^* = k/(k_m\sqrt{3}), \alpha = \mu_2/\mu_1, \\ x^* = k_m x \sqrt{3}, & L = k_m l \sqrt{3}, \quad f^*(x) = f(x)/U, \quad \lambda = 1/\sigma. \end{cases}$$

The maximum dimensionless value of the growth constant in the Saffman-Taylor case is $\sigma_{ST} = 2/(3\sqrt{3}) \approx 0.38$.

2. The model of Gorell and Homsy

An intermediate region PS, containing a polymer solute is considered in a model given by Gorell and Homsy, [2]. A steady basic solution, with straight initial interfaces exists. The tree regions are moving by the water velocity far upstream. The linear stability of interfaces is governed by a Sturm-Liouville problem, with eigenvalues (the growth constant of perturbations) in the boundary conditions. The unknown viscosity in PS, denoted by μ , is a parameter used to improve the stability of the PS - oil interface.

On the interfaces we consider the Laplace's law; moreover, on the interface water-PS a continuous viscosity is considered.

The Sturm-Liouville problem governing the stability is (written without *):

$$\begin{cases} -(\mu f_x)_x + k^2 \mu f = \lambda k^2 \beta \mu_x \cdot f, & x \in (-L, 0) \\ f_x(0) = (\lambda a + b) \cdot f(0) \\ f_x(-L) = k \cdot f(-L), \\ a = k^2 \beta \{\alpha - \mu(0) - k^2(\alpha - 1)\}/\mu(0), \quad \alpha = \mu_2/\mu_1 \\ b = -k \cdot \alpha/\mu(0), \quad \beta = (\alpha + 1)/(\alpha - 1), \quad \lambda = 1/\sigma. \end{cases}$$

where L is the length of PS. In a mobil system of coordinates, moving with the velocity U of water far upstream, the intermediate region is contained in the domain $x \in [-L, 0]$ - see [1] and [3]. The total amount C of polymer is:

$$C = \int_{-L}^0 \mu(x) dx.$$

The problem is to find the optimal value of μ d PS, which gives us the smallest growth constant σ , and to obtain an improvement of stability, compared with the Saffman-Taylor case. In [3] a numerical algorithm is used to give an "optimal" exponential viscosity profile in PS, which gives us a growth constant less than the Saffman-Taylor value.

3. The estimation of the growth constant

Carasso and Pasa, [1], obtain formula for an optimal viscosity in PS, and an upper estimation for the growth constant, in terms of $\mu(0)$ - the limit value of the viscosity in PS on the PS - oil interface. The above Sturm-Liouville problem is discretized by using the finite-difference method. We consider the points $x_i = -ih$, $h = L/N$, $i = 0, 1, \dots, N$ and obtain (by using the Greschgorin's localization theorem):

$$\sigma \leq \text{Max}\{H(k), \frac{\beta \mu'_i}{\mu_i}\}, \quad i = 1, 2, \dots, (N - 1), \quad \mu_i = \mu(x_i),$$

$$H(k) = \frac{k\beta}{\alpha}[\alpha - \mu(0) - k^2(\alpha - 1)], \quad ' = d/dx.$$

The function $H(k)$ has a maximum value with respect to k :

$$\text{Max}_k\{H(k)\} = F(\mu(0)) = \frac{2\beta}{3\alpha} \frac{(\alpha - \mu(0))^{3/2}}{\sqrt{3(\alpha - 1)}}.$$

We consider the viscosity profile (1) in PS and obtain the estimation (2):

$$(\beta\mu'_i)/\mu_i \leq F(\mu(0)) \quad (1)$$

$$\sigma \leq \text{Max}\{F(\mu(0)), \frac{\beta\mu'_i}{\mu_i}\} = F(\mu(0)) < \frac{2}{3\sqrt{3}} \approx \sigma_{ST}, \quad (2)$$

if the following condition holds:

$$F(\mu(0)) < 2/(3\sqrt{3}). \quad (3)$$

We imposed in [1] some restrictions for $\mu(0)$. The profile (1) gives us:

$$\mu(x) \leq \exp\{(x + L) \cdot \frac{F(\mu(0))}{\beta}\} \quad (4)$$

and for $x = 0$ we obtain the condition:

$$\mu(0) \leq \exp\{L \frac{F(\mu(0))}{\beta}\}. \quad (5)$$

We integrated (4) and obtained a relation involving the total amount of polymer C , the injection length L and the limit value $\mu(0)$:

$$L \geq \frac{\beta}{F(\mu(0))} \cdot \ln\left\{\frac{CF(\mu(0))}{\beta} + 1\right\}. \quad (6)$$

We can solve the following problem: for a given C , we compute the corresponding $\mu(0)$, L and the growth constant. Then it is not possible to obtain in a direct manner a "prescribed" improvement in stability, compared with the Saffman-Taylor case.

4. A new optimal profile in terms of $\mu(0)$

In this paper we can solve the inverse of the above problem. The relation (5) is considered as a "restriction" for L , in terms of $\mu(0)$:

$$L \geq \frac{\beta}{F(\mu(0))} \cdot \ln\{\mu(0)\}. \quad (7)$$

Therefore this time we have only one restriction for $\mu(0)$: it lies between water and oil viscosities. The relation (7) is considered as a second restriction for L . The point is to find an estimation of C in terms of $\mu(0)$ such that (6) and (7) would be compatible. For this, we consider the relations:

$$\begin{aligned} C &\geq \beta\{\mu(0)\} - 1\}/F(\mu(0)) \iff CF(\mu(0))/\beta + 1 \geq \mu(0) \iff \\ &\frac{\beta}{F(\mu(0))} \ln\{CF(\mu(0))/\beta + 1\} \geq \frac{\beta}{F(\mu(0))} \cdot \ln(\mu(0)). \end{aligned} \quad (8)$$

We use the relations (6) and (8) to obtain

$$L \geq \frac{\beta}{F(\mu(0))} \cdot \ln\left\{\frac{CF(\mu(0))}{\beta} + 1\right\} \geq \frac{\beta}{F(\mu(0))} \cdot \ln\{\mu(0)\}. \quad (9)$$

We consider the viscosity profile (1). Then the relations (2), (6), (8) give us the growth constant σ , the injection length L and the total amount C of polymer C in terms of $\mu(0)$. The limit value $\mu(0)$ verifies only one restriction: it lies between the water and oil viscosities. The condition (3) gives us an improved stability, compared with the Saffman–Taylor case. We emphasize that $\sigma \rightarrow 0$ for $\mu(0) \rightarrow \alpha$. We must avoid the case $\mu(0) = \alpha$, because (1) gives us $\mu' = 0$ and we obtain a constant viscosity in PS.

5. Numerical results

The function $F(\mu(0))$ is plotted in Figure 1, for the viscosity profile (1) and $\alpha = 100$, $\mu(0) \in [1, 90]$. The optimal viscosity profile in PS and the values of L are given in Figure 2, also in terms of $\mu(0)$. The relations (2), (6), (8) give us:

$$\mu(0) = 40 \Rightarrow F(\mu(0)) = \sigma_G = 0.17, \quad L \geq 20.49, \quad C \geq 216;$$

$$\mu(0) = 80 \Rightarrow F(\mu(0)) = \sigma_G = 0.03, \quad L \geq 126, \quad C \geq 2373;$$

$$\mu(0) = 90 \Rightarrow F(\mu(0)) = \sigma_G = 0.01, \quad L \geq 368, \quad C \geq 7442.$$

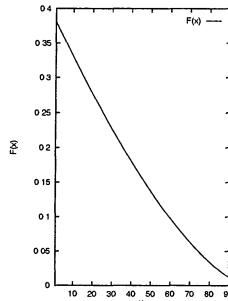
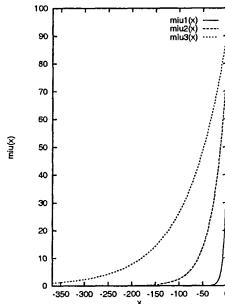


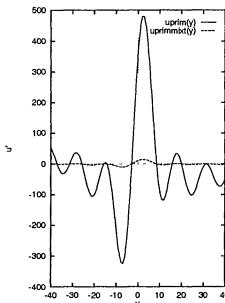
Figure 1. Maximal value $\sigma_G = F(\mu(0))$,

Figure 2. Optimal viscosity profiles: $\mu(0)=40, 80, 90$

In the Figure 3 are plotted the compared evolutions of the perturbations

$$\begin{aligned} & u'(x, y, t) = \\ & = f(x) \exp(\sigma \cdot t) \cdot \left\{ \sum_{n=2}^{n=9} [\cos(0.05 \cdot n \cdot y) + \sin(0.05 \cdot n \cdot y)] \right\} \end{aligned} \quad (10)$$

up to the time moment $t = 10$, for $\sigma = 0.38$ (without PS) and $\sigma = 0.03$. The initial amplitude is $f(0) \approx 1/10$.

Figure 3. Compared evolution of the perturbations (10), $t2=10$

References

- [1] Carasso, C., Paşa, G., An optimal viscosity profile in the secondary oil recovery, *Modélisation Mathématique et Analyse Numérique*, **32**, 1998, 211-221.
- [2] Gorell, S.B., Homsy, G.M., A theory of the optimal policy of oil recovery by the secondary displacement process, *SIAM J. Appl. Math.*, **43**, 1983, 79-98.
- [3] Saffman, P.G., Taylor, G., The penetration of a fluid in a porous medium or Hele-Shaw cell containing a more viscous liquid, *Proc.Roy.Soc., A* **245**, 1958, 312-329.

SINGULAR PERTURBATIONS OF HYPERBOLIC-PARABOLIC TYPE

Andrei Perjan

*Faculty of Mathematics and Informatics
Moldova State University
60 Mateevici Str., Chisinau 2009
Republic of Moldova
perjan@usm.md*

Abstract We study the behavior of solutions of the problem $\varepsilon u''(t) + u'(t) + Au(t) + Bu(t) = f(t)$, $u(0) = u_0$, $u'(0) = u_1$ in the Hilbert space H as $\varepsilon \rightarrow 0$, where A is linear, symmetric, continuous and strong positive operator and B is nonlinear lipschitzian operator.

Keywords: singular perturbations, hyperbolic equation, parabolic equation.

1. Statement of the problem

Let V and H be the real Hilbert spaces endowed with the norm $\|\cdot\|$ and $|\cdot|$ respectively such that $V \subset H \subset V'$, where every embedding is densely defined and continuous. By (\cdot, \cdot) we denote the duality between V and V' and also the scalar product in H . Let $A : V \rightarrow V'$ be a linear, continuous, symmetric operator and

$$(Au, u) \geq \omega \|u\|^2, \quad \forall u \in V, \omega > 0. \quad (1)$$

Let $B : H \rightarrow H$ be a nonlinear operator which satisfies the Lipschitz condition

$$|Bu - Bv| \leq L|u - v|, \quad \forall u, v \in H. \quad (2)$$

We shall study the behavior of the solutions of problem

$$\begin{cases} \varepsilon u''(t) + u'(t) + Au(t) + Bu(t) = f(t), \\ u(0) = u_0, u'(0) = u_1 \end{cases} \quad (P_\varepsilon)$$

as $\varepsilon \rightarrow 0$, where ε is a small positive parameter. Our aim is to show that $u \rightarrow v$ as $\varepsilon \rightarrow 0$, where v is the solution of the problem

$$\begin{cases} v'(t) + Av(t) + Bv(t) = f(t), \\ v(0) = u_0, \end{cases} \quad (P_0)$$

The main tool in our approach is the relation between the solutions of problem (P_ε) and the corresponding solutions of problem (P_0) in the linear case. This relation will be defined below by formula (3) and it was inspired by the work [1].

2. Notations

For $k \in \mathbb{N}$, $p \in [1, \infty)$ and $(a, b) \subset (-\infty, +\infty)$ by $W^{k,p}(a, b; H)$ denote the usual Sobolev spaces of vectorial distributions $W^{k,p}(a, b; H) = \{f \in D'(a, b; H); u^{(l)} \in L^p(a, b; H), l = 0, 1, \dots, k\}$ endowed with the norm

$$\|f\|_{W^{k,p}(a,b;H)} = \left(\sum_{l=0}^k \|f^{(l)}\|_{L^p(a,b;H)}^p \right)^{1/p}.$$

For $k \in \mathbb{N}$, $W^{k,\infty}(a, b; H)$ is the Banach space with the norm

$$\|f\|_{W^{k,\infty}(a,b;H)} = \max_{0 \leq l \leq k} \|f^{(l)}\|_{L^\infty(a,b;H)}$$

For $s \in \mathbb{R}$, $k \in \mathbb{N}$ and $p \in [1, \infty]$ define the following Banach spaces $W_s^{k,p}(a, b; H) = \{f : (a, b) \rightarrow H; e^{-st} f^{(l)} \in L^p(a, b; H), l = 0, 1, \dots, k\}$ with norms

$$\|f\|_{W_s^{k,p}(a,b;H)} = \|e^{-st} f^{(l)}(\cdot)\|_{W^{k,p}(a,b;H)}.$$

3. *A priori* estimates for solutions of the problem (P_ε)

At first we give the well-known existence theorems for the solutions of the problems (P_ε) and (P_0) [2].

Theorem A. *Suppose that $f \in W^{1,1}(0, T; H)$, $u_0 \in V$, $u_1 \in H$ and A and B satisfy conditions (1) and (2), then there exists a unique solution $u \in C(0, T; H) \cap L^\infty(0, T; V)$ of the problem (P_ε) which satisfies the following condition: $u' \in L^\infty(0, T; H) \cap C(0, T; V')$, $u'' \in L^\infty(0, T; V')$. If in addition $Au_0 \in H$, $u_1 \in V$, then $Au \in L^\infty(0, T; H)$, $u' \in L^\infty(0, T; V)$, $u'' \in L^\infty(0, T; H)$.*

Theorem B. *If $f \in W^{1,1}(0, T; H)$, $u_0 \in V$ and A and B satisfy the conditions (1), (2), then there exists a unique strong solution $v \in W^{1,\infty}(0, T; H)$ of the problem (P_0) and the estimates*

$$\begin{aligned} |v(t)| &\leq e^{Lt}(|u_0| + \int_0^t e^{-L\tau} (|B0| + |f(\tau)|) d\tau), \\ |v'(t)| &\leq Ce^{Lt}(|f(0)| + |\dot{A}u_0 + Bu_0| + \int_0^t e^{-L\tau} |f'(\tau)| d\tau), \end{aligned}$$

are true for $0 \leq t \leq T$.

Now we shall give the *a priori* estimates of the solutions of the problem (P_ε) . Denote by

$$\begin{aligned} E_1(u, t) &= \varepsilon|u'(t)| + |u(t)| + \left(\varepsilon(Au(t), u(t)) \right)^{1/2} + \\ &+ \left(\varepsilon \int_0^t |u'(\tau)|^2 d\tau \right)^{1/2} + \left(\int_0^t (Au(\tau), u(\tau)) d\tau \right)^{1/2}. \end{aligned}$$

Lemma 1. *Let $f \in W_{2L}^{1,1}(0, \infty; H)$, $u_0, u_1 \in H$, $Au_0, Au_1 \in H$ and A and B satisfy the conditions (1) and (2), then for every solution of the problem (P_ε) the following estimates*

$$\begin{aligned} E_1(u, t) &\leq Ce^{2Lt} \left(\varepsilon|u_1| + |u_0| + (\varepsilon(Au_0, u_0))^{1/2} + \|f\|_{L_{2L}^1(0, \infty; H)} \right), \quad t \geq 0, \\ E_1(u', t) &\leq Ce^{2Lt} \left(|f(0)| + |u_1| + |Au_0| + (\varepsilon(Au_1, u_1))^{1/2} + \right. \\ &\quad \left. + \|f'\|_{L_{2L}^1(0, \infty; H)} \right), \quad t \geq 0, \end{aligned}$$

are true.

4. The limit of solutions of the problem (P_ε) as $\varepsilon \rightarrow 0$

The studying of behavior of solutions of the problem (P_ε) as $\varepsilon \rightarrow 0$ is based on two key points. The first is getting the estimates of solutions which was obtained in Lemma 1 and the second is the relation (3) between the solutions of the problems (P_ε) and (P_0) . To establish the relation (3), at first we give some properties of the kernel $K(t, \tau)$ of transformation which realises this connection in the linear case, i. e. in the case when $B = 0$.

For $\varepsilon > 0$ denote

$$K(t, \tau) = \frac{1}{2\sqrt{\pi\varepsilon}} \left(K_1(t, \tau) + 3K_2(t, \tau) - 2K_3(t, \tau) \right),$$

where

$$\begin{aligned} K_1(t, \tau) &= e^{\frac{3t-2\tau}{4\varepsilon}} \lambda \left(\frac{2t-\tau}{2\sqrt{\varepsilon t}} \right), \\ K_2(t, \tau) &= e^{\frac{3t+6\tau}{4\varepsilon}} \lambda \left(\frac{2t+\tau}{2\sqrt{\varepsilon t}} \right), \\ K_3(t, \tau) &= e^{\frac{\tau}{\varepsilon}} \lambda \left(\frac{t+\tau}{2\sqrt{\varepsilon t}} \right), \end{aligned}$$

and $\lambda(s) = \int_s^\infty e^{-\eta^2} d\eta$.

Lemma 2. *The function $K(t, \tau)$ possesses the following properties:*

- (i) $K \in C(\bar{R}_+ \times \bar{R}_+) \cap C^2(R_+ \times R_+);$
- (ii) $K_t(t, \tau) = \varepsilon K_{\tau\tau}(t, \tau) - K_\tau(t, \tau), t > 0, \tau > 0;$
- (iii) $\varepsilon K_\tau(t, 0) - K(t, 0) = 0, t \geq 0;$
- (iv) $K(0, \tau) = \frac{1}{2\varepsilon} e^{-\frac{\tau}{2\varepsilon}}, \tau \geq 0;$
- (v) *For every fixed $t \geq 0$ there exist constants $C_1(t, \varepsilon) > 0, C_2(t, \varepsilon) > 0$ such that*

$$\begin{aligned} |K(t, \tau)| &\leq C_1(t, \varepsilon) e^{-C_2(t)\tau/\varepsilon}, & \tau > 0; \\ |K_t(t, \tau)| &\leq C_1(t, \varepsilon) e^{-C_2(t)\tau/\varepsilon}, & \tau > 0; \\ |K_\tau(t, \tau)| &\leq C_1(t, \varepsilon) e^{-C_2(t)\tau/\varepsilon}, & \tau > 0; \\ |K_{\tau\tau}(t, \tau)| &\leq C_1(t, \varepsilon) e^{-C_2(t)\tau/\varepsilon}, & \tau > 0; \end{aligned}$$

- (vi) $K(t, \tau) > 0, t \geq 0, \tau \geq 0;$
- (vii) *For every $C \in \mathbb{R}$ and every $\varphi : [0, \infty) \rightarrow H$ continuous on $[0, \infty)$ and $\varphi \in L_C^1(0, \infty; H)$ the relation*

$$\lim_{t \rightarrow 0} \int_0^\infty K(t, \tau) \varphi(\tau) d\tau = \int_0^\infty e^{-\tau} \varphi(2\varepsilon\tau) d\tau,$$

in H is valid;

- (viii) $\int_0^\infty K(t, \tau) d\tau = 1, t \geq 0;$
- (ix) *Let $\rho \in C^1[0, \infty), \rho$ and ρ' be increasing functions and $|\rho(t)| \leq M e^{ct}, |\rho'(t)| \leq M e^{ct}, t \in [0, \infty)$, then there exist positive constants C_1 and C_2 such that*

$$\int_0^\infty K(t, \tau) |\rho(t) - \rho(\tau)| d\tau \leq C_1 \sqrt{\varepsilon} e^{C_2 t}, t > 0;$$

- (x) *Let $C \geq 0$ and $f \in W_C^{1,\infty}(0, \infty : H)$, then*

$$|f(t) - \int_0^\infty K(t, \tau) f(\tau) d\tau|_H \leq C_1 \sqrt{\varepsilon} e^{C_2 t}, t \geq 0.$$

Theorem 1. *Let $A : D(A) \subset H \rightarrow H$ be a linear closed operator, $f \in L_C^\infty(0, \infty; H)$ (with real C). If u is a solution of the problem (P_ε)*

$(B = 0)$ such that $u \in W_C^{2,p}(0, \infty; H)$, $1 \leq p \leq \infty$, then the function v_0 , which is defined by

$$v_0(t) = \int_0^\infty K(t, \tau)u(\tau)d\tau \quad (3)$$

is the solution of the problem

$$\begin{cases} v'_0(t) + Av_0(t) = F(t, \varepsilon), \\ v_0(0) = \varphi_\varepsilon, \end{cases} \quad (P_0)$$

where

$$\begin{aligned} F(t, \varepsilon) &= \frac{1}{\sqrt{\pi}} \left(2e^{\frac{3t}{4\varepsilon}} \lambda \left(\sqrt{t/\varepsilon} \right) - \lambda \left(\frac{1}{2} \sqrt{t/\varepsilon} \right) \right) u_1 + \int_0^\infty K(t, \tau)f(\tau)d\tau, \\ \varphi_\varepsilon &= \int_0^\infty e^{-\tau} u(2\varepsilon\tau)d\tau. \end{aligned}$$

From Lemmas 1, 2 and from Theorem 1 the main result follows.

Theorem 2. Suppose that $f \in W_{2L}^{1,\infty}(0, \infty; H)$, $u_0, u_1 \in V$, $Au_0, Au_1 \in H$ and the operators A and B satisfy conditions (1) and (2). Then

$$|u(t) - v(t)| \leq C\sqrt{\varepsilon}, \quad 0 \leq t \leq T,$$

with C depending on $\|f\|_{W_{2L}^{1,\infty}(0, \infty; H)}$, $|u_0|$, $|u_1|$, $|Au_0|$, $|Au_1|$, ω and T .

Corollary 1. Let $B = 0$, operator A satisfy condition (1), $u_0, u_1, Au_0, Au_1 \in V$, $A^2u_0, A^2u_1 \in H$, $f \in W_{2L}^{2,\infty}(0, \infty; H)$ and

$$f(0) - u_1 - Au_0 = 0. \quad (4)$$

Then

$$|u'(t) - v'(t)| \leq C\varepsilon^{1/2}, \quad 0 \leq t \leq T,$$

$$\|u(t) - v(t)\| \leq C\varepsilon^{1/2}, \quad 0 \leq t \leq T.$$

It means that if condition (4) is satisfied, then the problem (P_0) is regularly perturbed, otherwise the problem (P_0) is singularly perturbed.

References

- [1] M.M. Lavrentiev, K.G. Reznitscaia, B.G. Iahno. The inverse one-dimentional problems from mathematecal physics. "Nauka", Novosibirsk, 1982 (in Russian).
- [2] V.Barbu. Semigroups of nonlinear contractions in Banach spaces. Bucarest, Ed. Acad. Rom., 1974 (in Rumanian).

IMPROVED DYNAMIC PROPERTIES BY FEEDBACK FOR SYSTEMS WITH DELAY IN CONTROL

Dan Popescu and Vladimir Răsvan

Department of Automatic Control

University of Craiova

13 A.I.Cuza Street, 1100 Craiova, ROMANIA

dpopescu@automation.ucv.ro; vrasvan@automation.ucv.ro

Abstract It is considered a linear controlled system with delay in control and subject to constant exogeneous signals, for which the following problems are discussed: feedback stabilization and exact regulation by finite pole assignment using a linear control law deduced by applying finite assignment and Artstein transform for a system that is extended by integrators; the same extended system is then stabilized subject to a quadratic performance index. The theorems are illustrated by simulation results. In simulation there are considered other effects: saturation of the actuators and intelligent integrators.

Keywords: Systems with delay in control, Stabilization, Linear Quadratic Problem.

AMS (MOS) Subject Classification: 93D15, 34K06, 34K35, 49K25.

1. Motivation and state of the art

It has been pointed out that systems with input delays are of interest to control theorists and practitioners for various reasons. They originate from the simplest model of process control which assigns to the controlled plant a transfer function of the form $H(s)e^{-\tau s}$ with $H(s)$ a strictly proper rational function. Systems arising from such transfer functions are of special type: their state space is finite dimensional but either the input or the output operators are defined on infinite dimensional extensions and are unbounded. Throughout several decades it has been established that such systems have "a finite dimensional flavor". Indeed, there exists a remark of V.M. Popov (1960) which tells that problems as stability and optimality in the framework of hyperstability (or dissipativity/passivity as it is called now) may be solved for these

systems like for systems without delay; nevertheless Popov did not follow this line in his research. We may add to this the results on feedback stabilization due to Olbrot (1978), Manitius and Olbrot (1979), Watanabe and Ito (1981). The techniques of these papers are essentially finite dimensional; the opinion of Pandolfi (1981, 1989, 1990, 1991) was that those results were not obtained in a standard way since they were not deduced from an abstract theory. In order to sustain his ideas, Pandolfi re-introduced in the model the *propagation effects* taken into account by the introduction of the delay and make use of the theory of the singular control. On the other hand, the finite dimensional results seem legitimate if the transform introduced by Artstein (1982) is used. The papers of Tadmor (e.g. 1995, 1998) also support the idea that finite dimension is the most adequate framework for systems with input delays. The authors of the present paper also followed the line opened by the papers of Olbrot and Manitius by using their approach in order to solve a LQ problem; also the hybrid piecewise constant control suggested by the paper of Halanay and Răsvan (1977) has been applied in the same framework and further completed by the continuous and discrete-time version of Artstein transforms (Popescu and Răsvan, 2001a, 2001b). In this paper we shall continue this line of research by considering the standard Linear Structurally Stable Regulator Problem (e.g. Francis, 1977; Francis and Wonham, 1975). For simplicity we shall consider here only the case of exogeneous constant (step) signals, a single input (control) signal and a single controlled (regulated) output, as in the paper of Ionescu and Răsvan (1991).

2. Problem statement and some preliminary results

The controlled system with input delay is the following

$$\begin{aligned}\dot{x}_1(t) &= A_1x_1(t) + A_3x_2(t) + b_{10}u(t) + b_{11}u(t - \tau) \\ \dot{x}_2(t) &= 0 \\ y(t) &= C_1x_1(t) + \int_{-\tau}^0 C_1e^{-A_1(\tau+\theta)}b_{11}u(t + \theta)d\theta + C_2x_2(t) \\ z(t) &= q^*y(t)\end{aligned}\tag{1}$$

Some explanations are necessary. Since we follow the line of the paper (Ionescu and Răsvan, 1991) the constant exogeneous signal vector (which incorporates both references and disturbances) is modelled by a suitably chosen autonomous system of differential equations of dimension n_2 while the state vector x_1 has dimension n_1 . Having in mind the transformation suggested by Artstein (1982) the measured output

$y(t)$ has been augmented by a term which takes into account the "history" of the control signal; if one takes into account that within the abstract model the past of the control signal belongs to the state and the outputs are always linear functionals defined on the state space, the structure of $y(t)$ turns to be legitimate. The form of $z(t)$ is given by the so-called *readability assumption*: the controlled output is readable from the measured output (e.g. Francis and Wonham, 1975).

Introducing the new state vector w_1 by

$$w_1(t) = x_1(t) + \int_{-\tau}^0 e^{-A_1(\tau+\theta)} b_{11} u(t+\theta) d\theta \quad (2)$$

the following controlled system is obtained

$$\begin{aligned} \dot{w}_1(t) &= A_1 w_1(t) + A_3 x_2(t) + (b_{10} + e^{-A_1\tau} b_{11}) u(t) \\ \dot{x}_2(t) &= 0 \\ y(t) &= C_1 w_1(t) + C_2 x_2(t) \\ z(t) &= q^* y(t) \end{aligned} \quad (3)$$

We may state now the following basic assumptions:

i) the pair $(A_1, b_{10} + e^{-A_1\tau} b_{11})$ is stabilizable and the pair (C_1, A_1) is detectable;

ii) the matrix $\begin{pmatrix} A_1 & b_{10} + e^{-A_1\tau} b_{11} \\ q^* C_1 & 0 \end{pmatrix}$ is non-singular i.e. $\lambda = 0$ is not a transmission zero of the triple $(A_1, b_{10} + e^{-A_1\tau} b_{11}, q^* C_1)$.

For system (3) we may state, as in (Ionescu and Răsvan, 1991) the Linear Structurally Stable Regulator Problem (LSSRP): find a measured-error-activated linear compensator

$$\begin{aligned} \dot{x}_c(t) &= A_c x_c(t) + B_c y(t) \\ u(t) &= f_c^* x_c(t) + g_c^* y(t) \end{aligned} \quad (4)$$

in order that the resulting closed loop linear system

$$\begin{aligned} \dot{w}_1(t) &= (A_1 + (b_{10} + e^{-A_1\tau} b_{11}) g_c^* C_1) w_1(t) + (A_3 + \\ &\quad (b_{10} + e^{-A_1\tau} b_{11}) g_c^* C_2) x_{20} + (b_{10} + e^{-A_1\tau} b_{11}) f_c^* x_c(t) \\ \dot{x}_c(t) &= B_c C_1 w_1(t) + B_c C_2 x_{20} + A_c x_c(t) \end{aligned} \quad (5)$$

should be asymptotically (in fact exponentially) stable in the autonomous case (in the absence of the exogeneous signals i.e. for $x_{20} = 0$) and $\lim_{t \rightarrow \infty} z(t) = 0$ for any $x_{20} \neq 0$; also these properties should hold for any A_3 and for small parameter uncertainties of A_1 , b_{10} , b_{11} , τ . Assume that LSSRP has been solved for the transformed system. We want, by

applying the inverse transform to (2), to see the structure achieved for the compensator and the properties of the resulting closed loop system. The feedback system in the variables x_1 , x_2 , x_c is as follows

$$\begin{aligned}\dot{x}_1(t) &= A_1x_1(t) + A_3x_2(t) + b_{10}u(t) + b_{11}u(t-\tau) \\ \dot{x}_2(t) &= 0 \\ \dot{x}_c(t) &= B_cC_1x_1(t) + B_cC_2x_2(t) + A_cx_c(t) + \\ &\quad B_cC_1 \int_{-\tau}^0 e^{-A_1(\tau+\theta)} b_{11}u(t+\theta) d\theta \\ u(t) &= q_c^*(C_1x_1(t) + C_2x_2(t)) + f_c^*x_c(t) + \\ &\quad g_c^*C_1 \int_{-\tau}^0 e^{-A_1(\tau+\theta)} b_{11}u(t+\theta) d\theta\end{aligned}\tag{6}$$

As a mathematical object system (6) is legitimate: let us substitute $u(t)$ from its integral equation in the first equation of (6) thus obtaining

$$\begin{aligned}\dot{x}_1(t) &= (A_1 + b_{10}g_c^*C_1)x_1(t) + (A_3 + b_{10}g_c^*C_2)x_2(t) + b_{10}f_c^*x_c(t) + \\ &\quad b_{11}u(t-\tau) + b_{10}g_c^*C_1 \int_{-\tau}^0 e^{-A_1(\tau+\theta)} b_{11}u(t+\theta) d\theta \\ \dot{x}_2(t) &= 0 \\ \dot{x}_c(t) &= B_c(C_1x_1(t) + C_2x_2(t)) + A_cx_c(t) + \\ &\quad B_cC_1 \int_{-\tau}^0 e^{-A_1(\tau+\theta)} b_{11}u(t+\theta) d\theta \\ u(t) &= q_c^*(C_1x_1(t) + C_2x_2(t)) + f_c^*x_c(t) + \\ &\quad g_c^*C_1 \int_{-\tau}^0 e^{-A_1(\tau+\theta)} b_{11}u(t+\theta) d\theta\end{aligned}\tag{7}$$

The solution of (7) may be constructed by steps, each step having two substeps, as follows: given $x_1(0) = x_{10}$, $x_2(0) = x_{20}$, $x_c(0) = x_{c0}$ and $u(\theta) = u^0(\theta)$, $-\tau \leq \theta < 0$ we can construct from the first three equations $x_1(t)$, $x_2(t)$ ($\equiv x_{20}$) and $x_c(t)$ on $[0, \tau]$; then we use the last equation to obtain $u(t)$ on the same interval $[0, \tau]$ and the first step is accomplished. We repeat this procedure on intervals $[k\tau, (k+1)\tau]$ with k a positive integer.

In a similar way we can discuss the internal stability property. Let $x_{20} = 0$ what gives $x_2(t) \equiv 0$. We know that (5) is exponentially stable hence there exist $\beta_1 > 0$, $\alpha > 0$ such that

$$|w_1(t)| + |x_c(t)| \leq \beta_1 e^{-\alpha t} (|w_1(0)| + |x_c(0)|) \tag{8}$$

We deduce from (2)

$$|x_1(t)| \leq |w_1(t)| + \int_{-\tau}^0 e^{\gamma(\tau+\theta)} |b_{11}| |u(t+\theta)| d\theta$$

On the other hand $w_1(0) = x_{10} + \int_{-\tau}^0 e^{-A_1(\tau+\theta)} b_{11} u^0(\theta) d\theta$ and, therefore

$$|w_1(0)| \leq |x_{10}| + \beta_2 \|u^0\|, \quad \|u^0\| = \left(\int_{-\tau}^0 |u^0(\theta)|^2 d\theta \right)^{1/2}$$

From (3) and (4) it follows that

$$|x_1(t)| \leq \left[\beta_1 + \beta_3 e^{\gamma\tau} \int_{-\tau}^0 e^{(\gamma-\alpha)\theta} d\theta \right] e^{-\alpha t} (|x_{10}| + \beta_2 \|u^0\| + |x_{c0}|)$$

what shows exponential stability of (6) (or (7)) in the autonomous case.

The regulator property $\lim_{t \rightarrow \infty} z(t) = 0$ for (6) - with $z(t)$ defined in (1) - follows at once from the regulator property of the finite dimensional closed loop system (5) with respect to the transformed controlled output $z(t)$. In this way we have proved the following preliminary result.

Theorem 1. Consider the controlled system (1) and the transformed system (3) via the change of coordinates (2). If (4) defines a regulating compensator for (3), solving LSSRP (Linear Structurally Stable Regulator Problem) in the sense that the autonomous system (with $x_2(t) \equiv 0$) is exponentially stable and $\lim_{t \rightarrow \infty} z(t) = 0$ for any $x_2(0) \neq 0$, the properties holding for any A_3 and for small parameter uncertainties of A_1 , b_{10} , b_{11} , $\tau > 0$, then the compensator defined by the last two equations of (6) solves LSSRP for (1).

3. Compensator synthesis

We shall recall here some results from (Ionescu and Răsvan, 1991) on LSSRP for the case of constant exogeneous signals. The structure of compensator (4) is as follows

$$A_c = \begin{pmatrix} A_w & a_a \\ 0 & 0 \end{pmatrix}, \quad B_c = \begin{pmatrix} B_w \\ q^* \end{pmatrix}, \quad f_c^* = (f_w^* \ f_a), \quad g_c^* \quad (9)$$

where $(A_w, a_a, B_w, f_w^*, f_a, g_c^*)$ defines a stabilizing compensator for the extended system

$$\begin{aligned} \dot{w}_e(t) &= A_e w_e(t) + b_e u(t) \\ y_e(t) &= C_e w_e(t) \end{aligned} \quad (10)$$

where

$$A_e = \begin{pmatrix} A_1 & 0 \\ q^* C_1 & 0 \end{pmatrix}, \quad b_e = \begin{pmatrix} b_{10} + e^{-A_1\tau} b_{11} \\ 0 \end{pmatrix}, \quad C_e = \begin{pmatrix} C_1 & 0 \\ 0 & 1 \end{pmatrix}, \quad (11)$$

$$w_e = \begin{pmatrix} w_1 \\ x_a \end{pmatrix}, \quad y_e = \begin{pmatrix} y \\ x_a \end{pmatrix}$$

and the compensator is given by

$$\begin{aligned}\dot{w}(t) &= A_w w(t) + B_w y(t) + a_a x_a(t) \\ u(t) &= f_w^* w(t) + g_c^* y(t) + f_a x_a(t)\end{aligned}\tag{12}$$

Worth mentioning that due to the basic assumptions *i)* and *ii)* such a stabilizing compensator can always be constructed while the structure resulting from the separation theorem - state feedback + observer - is not compulsory. On the contrary the integrator occurs in a necessary way and must be always present in the structure of the compensator; moreover, for a structurally stable design with respect to A_1 , b_{10} , b_{11} , A_3 , τ , this integrator must be contained in the compensator even if the initial system contains an integrator (i.e. has a zero eigenvalue). From (9) and (12) the following compensator equations can be written

$$\begin{aligned}\dot{w}(t) &= A_w w(t) + a_a x_a(t) + B_w y(t) \\ \dot{x}_a(t) &= q^* y(t) \\ u(t) &= f_w^* w(t) + f_a x_a(t) + g_c^* y(t)\end{aligned}\tag{13}$$

This compensator is implemented in (1) - with y defined as in (1) - in order to define the feedback system which is exponentially stable, has the regulator property for constant exogeneous signals with respect to the regulated output $z(t)$ and both these properties are robust i.e. structurally stable.

4. Optimal stabilization with respect to a quadratic performance index

We shall turn back to system (1) and associate it the simplest quadratic performance index

$$J(u) = \int_0^\infty [z^2(t) + \kappa u^2(t)] dt\tag{14}$$

where $\kappa > 0$, the index being defined on admissible pairs. Remark the overall system (1) has zero eigenvalues and is not completely controllable. We shall use some early results of V.M. Popov (1967) in order to obtain a state feedback law that ensures stabilization of the resulting closed loop system and a minimal value of (14).

Performing again the transformation (2) we obtain (3) which is finite dimensional and associate it to (14). In the following we shall check the assumptions required by the positivity theory of Popov for the uncontrollable case. In order to simplify the exposition we shall strengthen assumption *i)* of Section 2 as follows:

i') the pair $(A_1, b_{10} + e^{-A_1\tau}b_{11})$ is controllable and the pair (q^*C_1, A_1) is observable.

The characteristic function of the system, which coincides with the one of the controllable part (with the state vector w_1) is as follows

$$\chi(\lambda, \sigma) = \kappa + b_1^*(\lambda I - A_1^*)^{-1} C_1^* q q^* C_1 (\sigma I - A_1)^{-1} b_1 \quad (15)$$

where $b_1 = b_{10} + e^{-A_1\tau}b_{11}$; the characteristic polynomial is

$$\pi(\lambda, \sigma) = \det(\lambda I - A_1^*) \det(\sigma I - A_1) \chi(\lambda, \sigma) \quad (16)$$

We shall discuss first the assumptions $P_1 - P_3$ from the cited reference (Popov, 1967) since they may be considered as *non-degeneracy conditions*. We shall have

$$\pi(-i\omega, i\omega) = |\det(i\omega I - A_1)|^2 \left(\kappa + \left| q^* C_1 (i\omega I - A_1)^{-1} b_1 \right|^2 \right)$$

Since $\kappa > 0$, $\pi(-i\omega, i\omega)$ is not identically 0 hence $\pi(-\sigma, \sigma)$ is not identically 0 at least on $i\Re$; therefore P_1 holds. Moreover $\pi(-i\omega, i\omega) > 0$ for all ω hence P'_1 holds. Since A_{22} - the matrix defining the dynamics of x_2 - is the zero matrix, its only significant eigenvalue is 0 (with some multiplicity) P_2 holds because $\pi(0, 0) > 0$ (from P'_2 , in fact); obviously P_3 does not hold and we have to take this into account. In any case, since $\pi(-i\omega, i\omega) > 0$ and $\kappa > 0$ we deduce that $\chi(-i\omega, i\omega) > 0$. Using Theorem 1 of (Popov, 1967) we deduce existence of the polynomial $\psi(\sigma)$ with no roots in the closed RHP such that

$$\pi(-\sigma, \sigma) = \overline{\psi(-\bar{\sigma})}\psi(\sigma)$$

and also of the scalar γ , vectors p_1, p_2 and matrix N such that

$$\chi(-\sigma, \sigma) = \overline{\varphi(-\bar{\sigma})}\varphi(\sigma)$$

$$\varphi(\sigma) = \gamma + p_1^*(\sigma I - A_1)^{-1} b_1, \quad \varphi(\sigma) \det(\sigma I - A_1) \equiv \psi(\sigma)$$

$$\sqrt{\kappa} = \gamma, \quad N_{11} b_1 = p_1 \gamma$$

$$C_1^* q q^* C_1 + N_{11} A_1 + A_1^* N_{11} = p_1 p_1^*$$

$$N_{12}^* b_1 = p_2 \gamma, \quad A_3^* N_{11} + N_{12}^* A_1 + C_2^* q q^* C_1 = p_2 p_1^*$$

$$N_{12}^* A_3 + A_3^* N_{12} + C_2^* q q^* C_2 = p_2 p_2^*$$

where $N = \begin{pmatrix} N_{11} & N_{12} \\ N_{12}^* & N_{22} \end{pmatrix}$.

Here γ is computed in a straightforward way, p_1 follows from the factorization and N_{11} from a Liapunov equation. It may be shown by

following the lines of (Popov, 1967) that N_{12}^* and p_2 can be obtained from the above equations and that under these circumstances the last equation is automatically satisfied. Since N_{22} does not enter in these equations it might be chosen arbitrarily e.g. $N_{22} = 0$.

If we consider the integral

$$\eta(0, t_1) = \int_0^{t_1} [z^2(t) + \kappa u^2(t)] dt \quad (17)$$

then we may write, following (Popov, 1967)

$$\begin{aligned} \eta(0, t_1) &= -(w_1^* N_{11} w_1)|_0^{t_1} - (w_1^* N_{12} x_2)|_0^{t_1} - (x_2^* N_{12}^* w_1)|_0^{t_1} + \\ &\quad \int_0^{t_1} |\gamma u(t) + p_1^* w_1(t) + p_2^* x_2(t)|^2 dt \end{aligned} \quad (18)$$

We choose now

$$u(t) = -\frac{1}{\gamma}(p_1^* w_1(t) + p_2^* x_2(t)) \quad (19)$$

and system (3) becomes

$$\begin{aligned} \dot{w}_1(t) &= \left(A_1 - \frac{1}{\gamma} b_1 p_1^* \right) w_1(t) + \left(A_3 - \frac{1}{\gamma} b_1 p_2^* \right) x_2(t) \\ \dot{x}_2(t) &= 0 \end{aligned} \quad (20)$$

having bounded solutions. With the arguments of (Popov, 1967) convergence of the integral (14) along the solution defined by (19) is obtained. If we come back to the initial variables, the following dynamic feedback law is obtained:

$$u(t) + \frac{1}{\gamma} p_1^* \int_{-\tau}^0 e^{-A_1(\tau+\theta)} b_{11} u(t+\theta) d\theta = -\frac{1}{\gamma} (p_1^* x_1(t) + p_2^* x_2(t)) \quad (21)$$

and the optimal value of the index will depend on the initial history of u on $(-\tau, 0)$ - see also (Răsvan and Popescu, 2001a, 2001b).

5. An example. Simulation results

Let consider a system with input delay and constant disturbance described by

$$\dot{x}(t) = Ax(t) + B_1 u(t-\tau) + \nu_0 \quad (22)$$

where

$$A = \begin{pmatrix} 0 & 1 \\ -\omega_0^2 & 0 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

The controlled variable is

$$z(t) = Dx(t) \quad (23)$$

where $D = (1 \ 0)$.

In the design of high-accuracy regulators and tracking systems, it is necessary to eliminate completely the effect of offset errors caused by constant disturbances. This can be done by the application of integral control action. The destabilizing effect of the integral control, can be counteracted by the appropriate state feedback action, so that one can eventually achieve a satisfactory transient response, as well as the desired zero steady-state error.

Adjoin to the system variables the integral state" v , defined by

$$\dot{v}(t) = z(t) \quad (24)$$

with $v(0)$ given. Denoting the extended state vector by \bar{x}

$$\bar{x} = \begin{pmatrix} x \\ v \end{pmatrix} \quad (25)$$

the equations (22)–(24) can be written as follows

$$\begin{aligned} \dot{\bar{x}}(t) &= \bar{A}\bar{x}(t) + \bar{B}_1 u(t - \tau) + \bar{E}\nu_0 \\ z(t) &= \bar{D}\bar{x}(t) \end{aligned} \quad (26)$$

where

$$\bar{A} = \begin{pmatrix} 0 & 1 & 0 \\ -\omega_0^2 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \bar{B}_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \bar{E} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \bar{D} = (1 \ 0 \ 0)$$

It is easy to show that the (\bar{A}, \bar{B}_1) pair is controllable. Choosing a performance index of the form

$$J(u) = \int_0^\infty [\bar{x}^T(t)Q\bar{x}(t) + u^T(t)u(t)] dt \quad (27)$$

the solution of the optimization problem is determined

$$u(t) = -\bar{B}_1^T P e^{\bar{A}\tau} \left[\bar{x}(t) + \int_{-\tau}^0 e^{-\bar{A}(\tau+\theta)} \bar{B}_1 u(t+\theta) d\theta \right] \quad (28)$$

where P is the positive-definite solution of the matrix Riccati equation

$$Q + \bar{A}^T P + P \bar{A} - P \bar{B}_1 \bar{B}_1^T P = 0 \quad (29)$$

Then, the following piecewise constant control is applied to the system (22)

$$u(t) = -\bar{B}_1^T P e^{\bar{A}\tau} \left[\bar{x}_k + \sum_{-N}^{-1} \bar{A}(\delta)^{-(N+j+1)} \bar{B}_1(\delta) u_{k+j} \right], \quad (30)$$

$k\delta \leq t < (k+1)\delta$, where

$$\bar{\mathbf{A}}(\delta) = \begin{pmatrix} \cos(\omega_0\delta) & \frac{\sin(\omega_0\delta)}{\omega_0} & 0 \\ -\omega_0 \sin(\omega_0\delta) & \cos(\omega_0\delta) & 0 \\ \frac{\sin(\omega_0\delta)}{\omega_0} & \frac{1-\cos(\omega_0\delta)}{\omega_0^2} & 1 \end{pmatrix}, \quad \bar{\mathbf{B}}_1(\delta) = \begin{pmatrix} \frac{1-\cos(\omega_0\delta)}{\omega_0^3} \\ \frac{\sin(\omega_0\delta)}{\omega_0} \\ \frac{\omega_0\delta - \sin(\omega_0\delta)}{\omega_0^3} \end{pmatrix},$$

$$\delta = \tau/N$$

For simulation purposes, we choose $\omega_0 = \pi/2$ (rad/s), $\tau = 1$ (s), $N = 10$, $Q = \rho I$, $\rho = 10$. The simulation results obtained for a step disturbance v_0 , considering the actuator saturation for $|u(t)| > 2$ and "intelligent" integrators confirm the performance and robustness of the designed controller.

6. Concluding remarks

The application of the Artstein transform which reduces systems with delay in control to delayless systems allows to construct a finite dimensional theory for such systems. Nevertheless this is possible provided suitable output signals and integral indices are chosen.

References

- [1] Artstein, Z.(1982). Linear Systems with Delayed Controls: a Reduction, *IEEE Trans. Aut. Control*, vol.27, no.4, pp.869-879.
- [2] Drăgan, V. and A. Halanay(1999). *Stabilization of Linear Systems*, Birkhäuser Verlag, Boston.
- [3] Francis, B.(1977). The linear multivariable regulator problem, *SIAM J. Contr. Optim.*, vol.15, pp.486-505.
- [4] Francis, B. and W.M. Wonham(1975). The internal model principle for linear multivariable regulators, *Appl. Math. Optim.*, vol.2, no.2, pp.170-194.
- [5] Halanay, A. and Vl. Răsvan(1977). General Theory of Linear Hybrid Control, *Int. J. Control*, vol.20, no.4, pp.621-634.
- [6] Ionescu, V. and Vl. Răsvan(1991). The nonlinear regulator problem for constant signals, *Kybernetika*, vol.27, no.1, pp.12-22.
- [7] Krikelić, N.J.(1980). State feedback integral control with "intelligent" integrators, *Int. J. Control*, vol.32, no.3, pp.465-473.
- [8] Manitius, A. and A.W. Olbrot(1979). Finite spectrum assignment for systems with delays, *IEEE Trans. Aut. Control*, vol.AC-24, pp.541-553.
- [9] Olbrot, A.W.(1978). Stabilizability, detectability and spectrum assignment for linear systems with general time delays, *IEEE Trans. Aut. Control*, vol.AC-23, pp.887-890.
- [10] Pandolfi, L.(1981). A state space approach to control systems with delayed controls, In: *Proceedings of Conference on Functional Differential Equations and Related Topics (M. Kisielewicz, Ed.)*, pp.256-266, Zielona Gora, Poland.

- [11] Pandolfi, L.(1989). Dynamic stabilization of systems with input delays, In: *System Structure and Control: State Space and Polynomial Methods*, Prague.
- [12] Pandolfi, L.(1990). Generalized control systems, boundary control systems and delayed control systems, *MCSS*, vol.3, pp.165-181.
- [13] Pandolfi, L.(1991). Dynamic stabilization of systems with input delays, *Automatica*, vol.27, no.6, pp.1047-1050.
- [14] Popescu, D. and Vl. Răsvan(2001). Stabilization and Suboptimal Control of Systems with Input Delay, *The 11th Int. Symp. on Modeling, Simulation and Systems' Identification*, pp.259-266, Galati.
- [15] Popov, V.M.(1960). Stability criteria for systems containing non-univocal elements (in Romanian). In: *Probleme de Automatizare*, III, pp.143-151, Ed.Academiei, Bucharest.
- [16] Popov, V.M.(1967). Incompletely controllable positive systems and applications to optimisation and stability of automatic control systems. *Rev. Roum. Sci. Techn., Serie Electrot. et Energ.*, tome 12, no.3, pp.337-357.
- [17] Răsvan, Vl. and D. Popescu(2001a). Feedback Stabilization of Systems with Delays in Control, *Control Engineering and Applied Informatics*, vol.3, no.2, pp.62-66.
- [18] Răsvan, Vl. and D. Popescu(2001b). Control of systems with input delay by piecewise constant signals, *9th Medit. Conf. on Control and Automation*, Paper WM1-B/122, Dubrovnik, Croatia.
- [19] Tadmor, G.(1995). The Nehari problem in systems with distributed input delays is inherently finite dimensional, *Syst. and Contr. Let.*, vol.26, no.1, pp.11-16.
- [20] Tadmor, G.(1998). Robust control of systems with a single input lag, In: *Stability and Control of Time Lag Systems-LNCIS*, no.228, pp.259-282, Springer Verlag.
- [21] Watanabe, K. and M. Ito(1981). An observer for linear feedback control laws of multivariable systems with multiple delays in controls and output, *Syst. and Contr. Let.*, vol.1, no.1, pp.54-59.

GENERAL CONNECTIONS BETWEEN STRONG OPTIMIZATION AND PARETO EFFICIENCY

Vasile Postolica

Bacău State University

Department of Mathematical Sciences

vpostolica@ub.ro

Abstract In this research paper we present some connections between the Strong Optimization and the Approximate Pareto type efficiency in the usual Vector Optimization, first in Ordered Vector Spaces by the natural convex cones and, then, in Hausdorff Locally Convex Spaces. This fact may be very useful for possible scalarization methods.

Keywords: Pareto efficiency, nuclear (supernormal) cone, full nuclear cone.

1. Introduction

One of the direction concerning the scientific research in Vector Optimization is the study of the existence for Pareto type efficient points (see, for instance, the recent results, comments and examples in [1], [3]–[7], [10]–[14], [16]– [18]). In this context, it can be immediately remarked very clearly that an important mathematical tool, imposed especially by the implications and applications in the field of this kind of optimality dedicated to the efficiency, remains the concept of nuclear (supernormal) cone introduced by Isac G. [4], published in [5] and developed with significant applications until now in [5]–[8], [10]–[13], [17], [18].

The still unpublished research work [8] inspired us in this paper, having in mind the connections with the approximate Pareto efficiency, concept introduced here by us, that is, with its particular case represented by the usual Pareto efficiency. All the results, comments and examples are original.

2. The main results and related topics

Let X be a vector space ordered by a convex cone K , K_1 a non-void subset of K and A a non-empty subset of X . The following definition introduces a new concept of (approximate) Pareto type efficient points which leads to the well known notion of Pareto efficiency (in fact, the generalization in abstract spaces of the finite dimensional notion as we shall see in the next considerations).

Definition 2.1. We say that $a_0 \in A$ is a K_1 -Pareto (minimal) efficient point of A , in notation, $a_0 \in \text{eff}(A, K, K_1)$ (or $a_0 \in \text{MIN}_{K+K_1}(A)$) if it satisfies one of the following equivalent conditions:

- (i) $A \cap (a_0 - K - K_1) \subseteq a_0 + K + K_1$;
- (ii) $(K + K_1) \cap (a_0 - A) \subseteq -K - K_1$;

In a similar manner one defines the Pareto (maximal) efficient points by replacing $K + K_1$ with $-(K + K_1)$.

Remark 2.1. $a_0 \in \text{eff}(A, K, K_1)$ iff it is a fixed point for one of the following multifunctions $F_i : A \rightarrow A$, $i = \overline{1, 4}$ defined by

$$F_1(t) = \{a \in A : A \cap (a - K - K_1) \subseteq t + K + K_1\};$$

$$F_2(t) = \{a \in A : A \cap (t - K - K_1) \subseteq a + K + K_1\}.$$

Consequently, for the existence of the Pareto type efficient points one can apply appropriate fixed point theorems concerning the multifunctions (see, for instance, [1]).

Remark 2.2. in [2] it is shown that whenever $K_1 \subset K \setminus \{0\}$, the existence of this new type of efficient points for lower bounded sets characterizes the semi Archimedean ordered vector spaces and the regular ordered locally convex spaces.

Remark 2.3. When K is pointed, that is, $K \cap (-K) = \{0\}$ and $K_1 = \{0\}$, then, from Definition 2.1, we obtain the well known usual notion of Pareto (minimal, efficient, optimal or admissible) point, abbreviated $a_0 \in \text{eff}(A, K)$ (or $a_0 \in \text{MIN}_K(A)$), that is, satisfying the next equivalent properties:

- (i) $A \cap (a_0 - K) = \{a_0\}$;
- (ii) $A \cap (a_0 - K \setminus \{0\}) = \emptyset$;
- (iii) $K \cap (a_0 - A) = \{0\}$;
- (iv) $(K \setminus \{0\}) \cap (a_0 - A) = \emptyset$

and it is clear that for any $\varepsilon \in K \setminus \{0\}$, taking $K_1 = \{\varepsilon\}$, it follows that $a_0 \in \text{eff}(A, K, K_1)$ if and only if $A \cap (a_0 - \varepsilon - K) = \emptyset$. In all these cases, the set $\text{eff}((A, K, K_1))$ was denoted by $\varepsilon - \text{eff}(A, K)$ (or $\varepsilon - \text{MIN}_K(A)$ as in [2], [12] and [13]) and it is obvious that $\text{eff}(A, K) = \bigcap_{\varepsilon \in K \setminus \{0\}} [\varepsilon - \text{eff}(A, K)]$.

Remark 2.4. The following theorem offers the first important connection between the strong optimization and the (approximate) Pareto efficiency in the environment of ordered vector spaces, described initially on the previous Definition 2.1.

Theorem 2.1. *If we denote by $S(A, K, K_1) = \{a_1 \in A : A \subseteq a_1 + K + K_1\}$ and $S(A, K, K_1) \neq \emptyset$, then $S(A, K, K_1) = \text{eff}(A, K, K_1)$.*

Proof. Clearly, $S(A, K, K_1) \subseteq \text{eff}(A, K, K_1)$.

Indeed, if $a_0 \in S(A, K, K_1)$ and $a \in A \cap (a_0 - K - K_1)$ are arbitrary elements, then $a \in a_0 + K + K_1$, that is, $a_0 \in \text{eff}(A, K, K_1)$, by virtue of (i) in Definition 2.1. Suppose now that $\bar{a} \in S(A, K, K_1) \neq \emptyset$ and there exists $a_0 \in \text{eff}(A, K, K_1) \setminus S(A, K, K_1)$. From $\bar{a} \in S(A, K, K_1)$ it follows that $a_0 \in \bar{a} + K + K_1$, that is, $\bar{a} \in a_0 - K - K_1$, from which, since $\bar{a} \in A$ and $a_0 \in \text{eff}(A, K, K_1)$ we conclude that $\bar{a} \in a_0 + K + K_1$. Therefore, $A \subseteq \bar{a} + K + K_1 \subseteq a_0 + K + K_1$, in contradiction with $a_0 \notin S(A, K, K_1)$ as claimed.

Remark 2.5. The above theorem shows that, for any non-empty subset of an arbitrary vector space, the set of all strong minimal elements with respect to any convex cone through the agency of every non-void subset of it coincides with the corresponding set of Pareto (minimal) efficient points whenever there exists at least a strong minimal element, the result remaining obviously valid for the strong maximal elements and the Pareto maximal efficient points, respectively.

Using this result and our abstract construction given in [9] for the H -locally convex spaces introduced by Th. Precupanu in [15] as separated locally convex spaces with any seminorm satisfying the parallelogram law, we established in [12] that the only best simultaneous and vectorial approximation for each element in the direct sum of a (closed) linear subspace and its orthogonal with respect to a linear (continuous) operator between two H -locally convex spaces is its spline function. We also note that it is possible to have $S(A, K, K_1) = \emptyset$ and $\text{eff}(A, K, K_1) = A$. Thus, for example, if one considers $X = R^2$ endowed with the separated locally convex topology generated by the seminorms $p_1, p_2 : X \rightarrow R_+$, $p_1(x, y) = |x|$, $p_2(x, y) = |y|$, $K = R_+^2 = \{(x, y) \in R^2 : x, y \geq 0\}$, $K_1 = \{(0, 0)\}$ and $A = \{(\lambda, 1 - \lambda) : 0 \leq \lambda \leq 1\}$, then it is clear that $S(A, K, K_1) = \emptyset$ and $\text{eff}(A, K, K_1) = A$.

In all our further considerations we suppose that X is a Hausdorff locally convex space having the topology induced by family $P = \{p_\alpha : \alpha \in I\}$ of seminorms, ordered by a convex cone K and its topological dual space X^* . In this framework, the next theorem contains a significant criterion for the existence of the approximative Pareto (minimal) efficient points, in particular, for the usual Pareto (minimal) efficient points, taking into

account that the dual cone of K is defined by $K^* = \{x^* \in X^* : x^*(x) \geq 0, \forall x \in K\}$ and its attached polar cone is $K^0 = -K^*$. The version for the approximate Pareto (maximal) efficient points is straightforward.

Theorem 2.2. *If A is any non-empty subset of X and K_1 is every non-empty subset of K , then $a_0 \in \text{eff}(A, K, K_1)$ whenever for each $p_\alpha \in P$ and $\eta \in (0, 1)$ there exists x^* in the polar cone K^0 of K such that $p_\alpha(a_0 - a) \leq x^*(a_0 - a) + \eta, \forall a \in A$.*

Proof. We follow the general lines of the proof for Theorem 2.5 in [12].

Let us suppose that, under the above hypotheses,

$$(K + K_1) \cap (a_0 - A) \not\subseteq -(K + K_1),$$

that is, there exists $a \in A$ so that $a_0 - a \in K + K_1 \setminus (-K - K_1)$. Then, $a_0 - a \neq 0$ and, because X is separated in Hausdorff's sense, there exists $p_\alpha \in P$ such that $p_\alpha(a_0 - a) > 0$. On the other hand, there exists $n \in N^*$ sufficiently large with $p_\alpha(a_0 - a)/n \in (0, 1)$ and the relation given by the hypothesis of theorem leads to $p_\alpha(a_0 - a) \leq x^*(a_0 - a) + p_\alpha(a_0 - a)/n$ with $x^* \in K^0$ and $n \rightarrow \infty$, which implies that $p_\alpha(a_0 - a) \leq 0$, a contradiction and the proof is completed.

Remark 2.6. The above theorem represents an immediate extension of Precupanu's result given in Proposition 1.2 of [15]. In general, the converse of this theorem is not valid at least in (partially) ordered separated locally convex spaces as we can see from the example considered in Remark 2.5. Indeed, if one assumes the contrary in the corresponding, mathematical background, then, taking $\eta = \frac{1}{4}$ it follows that for each $\lambda_0 \in [0, 1]$ there exists $c_1, c_2 \leq 0$ such that $|\lambda_0 - \lambda| \leq (c_1 - c_2)(\lambda_0 - \lambda) + \frac{1}{4}$, $\forall \lambda \in [0, 1]$. Taking $\lambda_0 = \frac{1}{4}$ one obtains $|1 - 4\lambda| \leq (c_1 - c_2)(1 - 4\lambda) + 1$, $\forall \lambda \in [0, 1]$ which for $\lambda = 0$ implies that $c_2 \leq c_1$ and for $\lambda = \frac{1}{2}$ leads to $c_1 \leq c_2$, that is, $|1 - 4\lambda| \leq 1, \forall \lambda \in [0, 1]$, a contradiction.

The beginning and the considerations of Section 4 in [8] suggested us to consider for each function $\varphi : P \rightarrow K^* \setminus \{0\}$ the convex cone $K_\varphi = \{x \in X : p(x) \leq \varphi(p)(x), \forall p \in P\}$ and to extend Theorem 7 [8] in a more general context, which represents also a new link between strong optimization and the approximative vector optimization together with its usual particular variant, respectively.

Theorem 2.3. *If there exists $\varphi : P \rightarrow K^* \setminus \{0\}$ with $K \subseteq K_\varphi$, then*

$$\bigcup_{\substack{a \in A \\ \varphi : P \rightarrow K^* \setminus \{0\}}} S(A \cap (a - K - K_1), K_\varphi, \{0\}) \subseteq \text{eff}(A, K, K_1)$$

for any non-empty subset K_1 of K .

Proof. Let $a_1 \in S(A \cap (a_0 - K - K_1), K_\varphi, \{0\})$ for at least one elements $a_0 \in A$ and $\varphi : P \rightarrow K^* \setminus \{0\}$. Then, $a_1 \in A \cap (a_0 - K - K_1)$ and $A \cap (a_0 - K - K_1) - a_1 \subseteq K_\varphi$, that is, $p(a - a_1) \leq \varphi(p)(a - a_1), \forall a \in$

$A \cap (a_0 - K - K_1), p \in P$ which implies immediately that $p(a_1 - a) \leq -\varphi(p)(a_1 - a) + \eta, \forall a \in A \cap (a_0 - K - K_1), p \in P, \eta \in (0, 1)$ and, by virtue of Theorem 2.2 one obtains $a_1 \in eff(A \cap (a_0 - K - K_1), K, K_1)$. But $eff(A \cap (a_0 - K - K_1), K, K_1) \subseteq eff(A, K, K_1)$.

Indeed, for any $t \in eff(A \cap (a_0 - K - K_1), K, K_1)$ and $h \in A \cap (t - K - K_1)$ we have $h \in A \cap (a_0 - K - K_1) \cap (t - K - K_1) \subseteq t + K + K_1$ that is, $A \cap (t - K - K_1) \subseteq t + K + K_1$ and by point (i) of Definition 2.1 one obtains $t \in eff(A, K, K_1)$. This completes the proof.

Remark 2.7. The hypothesis $K \subseteq K_\varphi$ imposed upon the convex cone K is automatically satisfied whenever K is a supernormal (nuclear) cone. When K is any pointed convex cone, A is a nonempty subset of X and $a_0 \in eff(A, K)$, then, by virtue of (i) in Remark 2.3, it follows that $A \cap (a_0 - K) = \{a_0\}$, that is, $A \cap (a_0 - K) - a_0 = \{0\} \subset K_\varphi$. Hence, $a_0 \in S(A \cap (a_0 - K), K_\varphi, \{0\})$ for every mapping $\varphi : P : K^* \setminus \{0\}$ and the next corollary is valid.

Corollary 2.1. *For every non-empty subset A of any Hausdorff locally convex space ordered by an arbitrary, pointed convex cone K with its dual cone K^* we have*

$$eff(A, K) = \bigcup_{\substack{a \in A \\ \varphi : P \rightarrow K^* \setminus \{0\}}} S(A \cap (a - K), K_\varphi)$$

Remark 2.8. Clearly, the result concerns the possibilities of scalarization for the study of Pareto efficiency in separated locally convex spaces, as we can see also in the final comments of [8] for the particular cases of Hausdorff locally convex spaces ordered by closed, pointed and normal cones.

References

- [1] Luc D.T.: *Theory of Vector Optimization*, Springer-Verlag, 1989.
- [2] Cardinali T., Papalini, F.: Fixed point theorems for multifunctions in topological vector spaces. *Journal of Mathematical Analysis and Applications* **186**, 1994, 769–777.
- [3] Nemeth, A.B.: Between Pareto efficiency and Pareto ε -efficiency. *Optimization*, **20**, 5 (1989), 615–637.
- [4] Isac G.: Points critiques des systemes dynamiques. Cones nucleaires et optimum de Pareto. *Research Report*, Royal Military College of St. Jean, Quebec, Canada, 1981.
- [5] Isac G.: Sur l'existence de l'optimum de Pareto. *Riv. Mat. Univ. Parma* (4), **9** 1983, 303–325.
- [6] Isac G., Postolica V.: *The Best Approximation and Optimization in Locally Convex Spaces*. Verlag Peter Lang GmbH, Frankfurt am Main, Germany, 1993.
- [7] Isac G.: Pareto optimization in infinite dimensional spaces: the importance of nuclear cones. *Journ. of Math. Anal. and Appl.*, **182**, no. 2, 1994, 393–404.

- [8] Isac G.: On Pareto efficiency. A general constructive existence principle. Research Report (16 pp.). Department of Mathematics and Computer Science, Royal Military College of Canada, 1998.
- [9] Isac G., Bahya A.O.: Full nuclear cones associated to a normal cone. Application to Pareto efficiency. To appear in Applied Math. Letters, 2002.
- [10] Postolică V.: Spline functions in H -locally convex spaces. *An. St. Univ. "Al.I.Cuza" Iași*, Romania, 27, 1981, 333–338.
- [11] Postolică V.: New existence results for efficient points in locally convex spaces ordered by supernormal cones. *Journal of Global Optimization* 3, 1993, 233–242.
- [12] Postolică V.: Properties of Pareto set in locally convex spaces. *Optimization*, vol. 34, 1995, 223–229.
- [13] Postolică V.: Properties of efficient points sets and related topics. Research Report, at The Second International Conference on Multi-Objective Programming and Goal Programming, Torremolinos, Spain, May 16–18, 1996. Published in *Advances in Multiple Objective and Goal Programming*. Lecture Notes in Economics and Mathematical Systems 455, Rafael Cabalberro, Francisco Ruiz, Ralph E. Steuer (Eds), 1997, 201–209, Springer Verlag, Berlin.
- [14] Postolică V.: Efficiency and Choquet boundaries in separated locally convex spaces. (ubmited), 2001.
- [15] Precupanu Th.: Espaces lineaires à semi-normes hilbertiennes. *An. St. Univ. "Al.I.Cuza" Iași*, Romania, 15, 1969, 83–93.
- [16] Precupanu, Th.: *Scalar minimax properties in vectorial optimization*. International Series of Numerical Mathematics, Birkhauser Verlag Basel, vol. 107, 1992, 299–306.
- [17] Truong X.D.Ha: A note on a class of cones ensuring the existence of efficient points in bounded complete sets. *Optimization*, vol. 31, 1994, 141–152.
- [18] Truong X.D.Ha: On the existence of efficient points in locally convex spaces. *Journal of Global Optimization*, vol.4, 1994, 265–278.

GENERALIZED HO-KALMAN ALGORITHM FOR 2D CONTINUOUS DISCRETE LINEAR SYSTEMS

V. Prepeliță

Universitatea Politehnica București, Catedra Matematici I,

Splaiul Independenței 313, Sector 6, Cod 77206, București, Romania

vprepelita@pcnet.ro

Abstract A class of continuous-discrete time-variable linear control systems is considered, whose state space representation is a system of differential equations with respect to one variable and of difference equations with respect to the second one. The fundamental concepts of reachability and observability are analysed in this framework. In the case of time-invariant systems the structure of the transfer matrix is obtained and some properties of minimal realizations are emphasized. The connection between reachability, observability and minimality is established. An algorithm is proposed which provides a minimal realization for multi-input-multi-output systems. This method generalizes to 2D systems the celebrated Ho-Kalman algorithm.

MSC 1991: 93B15, 93B20, 93C35

Keywords. 2D continuous-discrete systems, controllability, observability, minimal realizations.

1. Introduction

Last years proved an increasing interest in the study of multivariable systems, determined by their wide range of applications in domains like image processing, geophysics and seismology, computer tomography etc.

One of the most important problems in this approach is that of the minimal state-space realization of nD transfer matrices. As it was emphasized in B. De Schutter's overview [1], the notion of minimality plays a powerful role in the analysis and design of nD digital signal processing. It has been proved that minimal state space realizations are possible only in special cases of rational $2D$ transfer functions with separable denominator, separable numerator, for the all-pole and all-zero systems or for the continued fraction expansions (see [2], [3] and [4]). The disadvantage

of the existing algorithms is their restriction to single-input-single-output (SISO) systems.

In this paper an algorithm is proposed which determines a minimal realization for separable $2D$ multi-input-multi-output (MIMO) systems. This method generalizes to $2D$ systems the celebrated Ho-Kalman algorithm (see [5]) which gave rise to many interesting developments (see [1] and [6]). A similar procedure was proposed in [7] in Kalman's module theoretic approach.

The basic concepts of controllability and observability are discussed in the general case of time-variable systems and suitable Gramians and matrices are defined. The connection between the controllability matrix, the observability matrix and the Hankel matrix of a $2D$ time-invariant system is emphasized and it is used to prove the minimality of the provided realization.

The above topics are related to a class of continuous-discrete linear control systems whose state space representation is a system of differential equations with respect to one variable and of difference equations with respect to the second one. This class corresponds to Attasi's two dimensional discrete-time linear systems (see [8]), and in the time-invariant case was studied in [9]. Similar hybrid systems (but of Roesser [10] and Fornasini-Marchesini [11] type) were studied in a series of papers due to Kaczorek (see [12], [13] and [14]). The study of $2D$ continuous-discrete systems is motivated by their applications in various domains like discrete linear repetitive processes (or multipass processes), in iterative learning control or in the study of linear systems with delays (see [15], [16] and [17]).

The proposed algorithm can also be used for MIMO separable $2D$ discrete-time linear systems or for MIMO $2D$ systems described by a class of hyperbolic partial differential equations.

2. Two-Dimensional Continuous-Discrete Linear Systems

The linear spaces $X=\mathbf{R}^n$, $U=\mathbf{R}^m$ and $Y=\mathbf{R}^p$ are called respectively the *state*, *input* and *output spaces* and $T=\mathbf{R} \times \mathbf{Z}$ is the time set.

Definition 2.1. A *two-dimensional continuous-discrete linear system* ($2Dcd$) is a quintuplet $\Sigma = (A_1(t, k), A_2(t, k), B(t, k), C(t, k), D(t, k)) \in \mathbf{R}^{n \times n} \times \mathbf{R}^{n \times n} \times \mathbf{R}^{n \times m} \times \mathbf{R}^{p \times n} \times \mathbf{R}^{p \times m}$ with $A_1(t, k)A_2(t, k)=A_2(t, k)A_1(t, k)$ $\forall (t, k) \in T$, where all matrices are continuous with respect to $t \in \mathbf{R}$ for any $k \in \mathbf{Z}$; the state space representation of Σ is given by the state and

output equations

$$\begin{aligned}\dot{x}(t, k+1) &= A_1(t, k+1)x(t, k+1) + A_2(t, k)\dot{x}(t, k) - \\ &\quad - A_1(t, k)A_2(t, k)x(t, k) + B(t, k)u(t, k)\end{aligned}\tag{2.1}$$

$$y(t, k) = C(t, k)x(t, k) + D(t, k)u(t, k)\tag{2.2}$$

where $\dot{x}(t, k) = \frac{\partial x}{\partial t}(t, k)$. The number n is called the *dimension* of the system Σ and it is denoted by $\dim\Sigma$.

All the following results remain valid in the more general context of matrices over spaces of functions of bounded variation or regulated functions with respect to t , in which case the Perron-Stieltjes integral is used (see [18], [19] and [20]).

The (continuous) fundamental matrix of $A_1(t, k)$ with respect to $t \in \mathbf{R}$ is denoted by $\Phi(t, t_0; k)$ for any fixed $k \in \mathbf{Z}$; therefore $\Phi(t, t_0; k)$ is the unique matrix solution of the system $\dot{Y}(t, k) = A_1(t, k)Y(t, k)$, $Y(t_0, k) = I$. If A_1 is a constant matrix, then $\Phi(t, t_0; k) = e^{A_1(t-t_0)}$.

The *discrete fundamental matrix* $F(t; k, k_0)$ of the matrix $A_2(t, k)$ is defined for any fixed $t \in \mathbf{R}$ by

$$F(t; k, k_0) = \begin{cases} A_2(t, k-1)A_2(t, k-2)\dots A_2(t, k_0) & \text{for } k > k_0 \\ I_n & \text{for } k = k_0. \end{cases}$$

$F(t; k, k_0)$ is the unique matrix solution of the system of difference equations $Y(t, k+1) = A_2(t, k)Y(t, k)$, $Y(t, k_0) = I$. If A_2 is a constant matrix, then $F(t; k, k_0) = A_2^{k-k_0}$.

Since $A_1(t, k)$ and $A_2(t, k)$ are commutative matrices for any $(t, k) \in T$, it results by Peano-Baker formula for Φ and by the definition of F that $\Phi(t, t_0; k)$ and $F(s; l, l_0)$ are commutative matrices for any $t, t_0, s \in \mathbf{R}$ and $k, l, l_0 \in \mathbf{Z}$.

Definition 2.2. A vector $x_0 \in X$ is said to be the *initial state* of Σ at the moment $(t_0, k_0) \in T$ if for any $(t, k) \in T$ with $(t, k) \geq (t_0, k_0)$ the following conditions hold:

$$x(t, k_0) = \Phi(t, t_0; k_0)x_0 \quad x(t_0, k) = F(t_0; k, k_0)x_0 \tag{2.3}.$$

For some $(t_0, k_0), (t, k) \in T$ with $(t_0, k_0) < (t, k)$ we denote by I the set $I = [t_0, t] \times [k_0, k]$, $I \subset T$. An *input function* or a *control* is a function $u : I \rightarrow U$ such that $u(\cdot, k)$ is continuous for any $k \in \mathbf{Z}$.

In [21] it was proved that the state of Σ at the moment $(t, k) \in T$ determined by the initial state x_0 and by the input function u is given by

$$\begin{aligned}x(t, k) &= \Phi(t, t_0; k)F(t_0; k, k_0)x_0 + \\ &+ \int_{t_0}^t \sum_{l=k_0}^{k-1} \Phi(t, s; k)F(s; k, l+1)B(s, l)u(s, l)ds.\end{aligned}\tag{2.4}$$

By replacing the state $x(t, k)$ given by (2.4) into the output equation (2.2) we obtain the input-output map of the system Σ :

$$\begin{aligned} y(t, k) = & C(t, k)\Phi(t, t_0; k)F(t_0; k, k_0)x_0 + \\ & + \int_{t_0}^t \sum_{l=k_0}^{k-1} C(t, k)\Phi(t, s; k)F(s; k, l+1)B(s, l)u(s, l)ds + D(t, k)u(t, k). \end{aligned} \quad (2.5)$$

Definition 2.3. A triplet $(t, k, x) \in \mathbf{R} \times \mathbf{Z} \times X$ is said to be a *phase* of Σ if x is the state of Σ at the moment (t, k) (i.e. if $x = x(t, k)$, where $x(t, k)$ is given by (2.4)).

By $(s, l) < (t, k)$ we mean $s \leq t, l \leq k$ and $(s, l) \neq (t, k)$.

Definition 2.4. A phase (t, k, x) of Σ is said to be *reachable* if there exist $(t_0, k_0) \in T$, $(t_0, k_0) < (t, k)$ and a control $u(\cdot, \cdot)$ which transfers the phase $(t_0, k_0, 0)$ to (t, k, x) .

A phase (t, k, x) is said to be *controllable* if there exist $(t_1, k_1) \in T$, $(t_1, k_1) > (t, k)$ and a control $u(\cdot, \cdot)$ which transfers the phase to (t, k, x) to $(t_1, k_1, 0)$.

If for some fixed $(t_0, k_0) \in T$, $(t_1, k_1) \in T$, $(t_0, k_0) < (t_1, k_1)$, every phase (t_1, k_1, x) $((t_0, k_0, x))$ is reachable (controllable) during the period $[t_0, t_1] \times [k_0, k_1] \subset \mathbf{R} \times \mathbf{Z}$, the system Σ is said to be *completely reachable* (*completely controllable*) on $[t_0, t_1] \times [k_0, k_1]$.

The symmetrical non-negative definite $n \times n$ matrix

$$R_\Sigma(t_0, t; k_0, k) = \quad (2.6)$$

$$\int_{t_0}^t \sum_{l=k_0}^{k-1} \Phi(t, s; k)F(s; k, l+1)B(s, l)B(s, l)^T F(s; k, l+1)^T \Phi(t, s; k)^T ds$$

is called the *reachability Gramian* of Σ .

In [21] it was proved:

Proposition 2.5. *The set of all states which are reachable on I is the linear space $\text{Im}R_\Sigma(t_0, t; k_0, k)$.*

Then we get

Theorem 2.6. *Σ is completely reachable on I if and only if*

$$\text{Rank}R_\Sigma(t_0, t; k_0, k) = n. \quad (2.7)$$

Proof. By Proposition 2.5, Σ is completely reachable on I iff $\text{Im}R_\Sigma(t_0, t; k_0, k) = \mathbf{R}^n$, condition equivalent to (2.7).

Definition 2.7. A phase (t_0, k_0, x) is said to be *unobservable* (*unobservable on I*) if for any control $u : I \rightarrow \mathbf{R}^m$ it provides the same output $y(s, l)$ for $(s, l) \geq (t, k)$ (for $(s, l) \in I$) as the phase $(t_0, k_0, 0)$.

The system Σ is said to be *completely observable* (*completely observable on I*) if there is no unobservable (unobservable on I) state $x \neq 0$.

In order to check whether a system Σ is completely observable we introduce the 2D *observability Gramian* of Σ denoted by $\mathcal{O}_\Sigma(t, t_0; k, k_0)$:

$$\begin{aligned} & \mathcal{O}_\Sigma(t, t_0; k, k_0) = \\ &= \int_{t_0}^t \sum_{l=k_0}^k F(t_0; l, k_0)^T \Phi(s, t_0; l)^T C(s, l)^T C(s, l) \Phi(s, t_0; l) F(t_0; l, k_0) ds. \end{aligned}$$

The following results are proved in [22]:

Proposition 2.8. *The set of states which are unobservable on I is the subspace $\text{Ker } \mathcal{O}_\Sigma(t, t_0; k, k_0)$.*

Theorem 2.9. *The system $\Sigma = (A_1(t, k), A_2(t, k), B(t, k), C(t, k), D(t, k))$ is completely observable on I if and only if*

$$\text{rank } \mathcal{O}_\Sigma(t, t_0; k, k_0) = n. \quad (2.8)$$

3. Time-Invariant 2Dcd Systems

The system $\Sigma = (A_1, A_2, B, C, D)$ is said to be *time invariant* if A_1, A_2, B, C and D are constant matrices. In this case the state formula (2.4) and the output formula (2.5) become

$$x(t, k) = e^{A_1(t-t_0)} A_2^{k-k_0} x_0 + \int_{t_0}^t \sum_{l=k_0}^{k-1} e^{A_1(t-s)} A_2^{k-l-1} B u(s, l) ds \quad (3.1)$$

$$y(t, k) = C e^{A_1 t} A_2^k x_0 + \int_0^t \sum_{l=0}^{k-1} C e^{A_1(t-s)} A_2^{k-l-1} B u(s, l) ds + D u(t, k). \quad (3.2)$$

Since Σ is time-invariant, we can consider the initial moment $(t_0, k_0) = (0, 0)$ and the time set $T = \mathbf{R}^+ \times \mathbf{Z}^+$.

We associate to Σ the *controllability matrix*

$$C_\Sigma = [B \ A_1 B \ \dots \ A_1^{n-1} B \ A_2 B \ A_1 A_2 B \ \dots \ A_1^{n-1} A_2 B \dots \ A_2^{n-1} B \ A_1 A_2^{n-1} B \ \dots \ A_1^{n-1} A_2^{n-1} B]. \quad (3.3)$$

From Theorem 2.6 we get (see [21])

Theorem 3.1. Σ is completely reachable if and only if

$$\text{rank } C_\Sigma = n. \quad (3.4)$$

Theorem 3.2. *The set of all reachable states of Σ is the smallest subspace of X which is (A_1, A_2) -invariant and contains the columns of B .*

The matrix

$$O_{\Sigma} = [C^T \ A_1^T C^T \ \dots \ (A_1^T)^{n-1} C^T \ A_2^T C^T \ A_1^T A_2^T C^T \ \dots \ (A_1^T)^{n-1} A_2^T C^T \ \dots \ (A_2^T)^{n-1} C^T \ A_1^T (A_2^T)^{n-1} C^T \ \dots \ (A_1^T)^{n-1} (A_2^T)^{n-1} C^T]^T \quad (3.5).$$

is called the *observability matrix* of the system Σ .

In [22] it was proved:

Theorem 3.3. *The system $\Sigma = (A_1, A_2, B, C, D)$ is completely observable if and only if*

$$\text{rank } O_{\Sigma} = n. \quad (3.6)$$

The following statement results from Theorems 3.1 and 3.3 by noticing that the 2D observability matrix O_{Σ} of the system $\Sigma = (A_1, A_2, B, C, D)$ coincides with the 2D controllability matrix C_{Σ^*} of the system $\Sigma^* = (A_1^T, A_2^T, C^T, B^T, D^T)$. We say that the system Σ^* is the *dual* of Σ .

Theorem 3.4. *The system Σ is completely observable if and only if its dual Σ^* is completely reachable.*

By duality we obtain (see [22]):

Theorem 3.5. *The set X_{uo} of all unobservable states of Σ is the greatest subspace of X which is (A_1, A_2) -invariant and is contained in $\text{Ker } C$.*

4. Transfer Matrix of 2Dcd Systems

Let us consider the time-invariant 2Dcd system $\Sigma = (A_1, A_2, B, C, D)$. The initial conditions (2.3) become (with $x_0 = x(0, 0)$)

$$x(t, 0) = e^{A_1 t} x_0, \forall t \in \mathbf{R}^+; x(0, k) = A_2^k x_0, \forall k \in \mathbf{Z}^+, \quad (4.1)$$

hence

$$\dot{x}(t, 0) = A_1 x(t, 0); x(0, k+1) = A_2 x(0, k). \quad (4.2)$$

We denote by $\tilde{x}(s, k)$ the Laplace Transform of the function $x(t, k)$ for $k \in \mathbf{Z}^+$ and by $X(s, z)$ the z -Transform of $\tilde{x}(s, k)$ for $s \in \mathbf{C}$. The Differentiation-of-the-original Theorem for the Laplace Transformation and Second-delay Theorem for z -Transformation give

$$s\tilde{x}(s, 0) - x(0, 0) = A_1 \tilde{x}(s, 0); zX(0, z) - zx(0, 0) = A_2 X(0, z). \quad (4.3)$$

By applying the Laplace Transformation to (2.1) we get $s\tilde{x}(t, k+1) - x(0, k+1) = A_1 \tilde{x}(s, k+1) + A_2(s\tilde{x}(s, k) - x(0, k)) - A_1 A_2 \tilde{x}(s, k) + B\tilde{u}(s, k)$,

equation in which we can reduce $x(0, k+1)$ by (4.2). Then we take the z-Transform of the reduced equation and we obtain $z(sX(s, z) - s\tilde{x}(s, 0)) = A_1z(X(s, z) - \tilde{x}(s, 0)) + A_2sX(s, z) - A_1A_2X(s, z) + BU(s, z)$ which gives by (4.3) $(zsI - A_1 - A_2 + A_1A_2)X(s, z) = BU(s, z) + zx_0$.

For $s \in \mathbf{C} \setminus \sigma(A_1), z \in \mathbf{C} \setminus \sigma(A_2)$ we have

$$X(s, z) = (sI - A_1)^{-1}(zI - A_2)^{-1}BU(s, z) + (sI - A_1)^{-1}(zI - A_2)^{-1}zx_0. \quad (4.4)$$

Both Laplace and z Transforms applied to (2.2) give $Y(s, z) = CX(s, z) + DU(s, z)$ and by replacing $X(s, z)$ (4.4) we obtain the input-output map of Σ in the frequency domain

$$Y(s, z) = \dots \quad (4.5)$$

$$[C(sI - A_1)^{-1}(zI - A_2)^{-1}B + D]U(s, z) + C(sI - A_1)^{-1}(zI - A_2)^{-1}zx_0.$$

Definition 4.1. The matrix

$$T_\Sigma(s, z) = C(sI - A_1)^{-1}(zI - A_2)^{-1}B + D \quad (4.6)$$

is called the *transfer matrix* of Σ .

Obviously, $T_\Sigma(s, z)$ is a $p \times m$ rational proper (in both variables s and z) matrix with separable denominator, since it has the form

$$T_\Sigma(s, z) = \frac{1}{\det(sI - A_1)\det(zI - A_2)}C(sI - A_1)^*(zI - A_2)^*B + D.$$

With $x_0 = 0$ in (4.5) we get

Proposition 4.2. *The null state response of the system $T_\Sigma(s, z)$ in the frequency domain has the form $Y(s, z) = T_\Sigma(s, z)U(s, z)$ where $T_\Sigma(s, z)$ is the transfer matrix of Σ .*

Let us denote by $PS(m, p)$ the set of $p \times m$ proper matrices with separable denominator and by $SPS(m, p)$ the set of strictly proper matrices in $PS(m, p)$.

Definition 4.3. Given $T(s, z) \in PS(m, p)$, a system $\Sigma = (A_1, A_2, B, C, D)$ is said to be a *realization* of Σ if $T(s, z) = T_\Sigma(s, z)$, that is if

$$T(s, z) = C(sI - A_1)^{-1}(zI - A_2)^{-1}B + D. \quad (4.7)$$

A realization Σ of $T(s, z)$ is *minimal* if $\dim \Sigma \leq \dim \tilde{\Sigma}$ for any realization $\tilde{\Sigma}$ of $T(s, z)$.

Since by (4.7) $D = \lim_{s \rightarrow \infty} T(s, z) = \lim_{z \rightarrow \infty} T(s, z)$, we can state the usual *realization problem*: given a strictly proper matrix $T(s, z)$ determine the quadruplet $\Sigma = (A_1, A_2, B, C)$ such that

$$T(s, z) = C(sI - A_1)^{-1}(zI - A_2)^{-1}B. \quad (4.8)$$

Now let us consider the Laurent series expansion of $T(s, z) \in SPS(p, m)$ about $s = \infty, z = \infty$

$$T(s, z) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} M_{i,j} s^{-i-1} z^{-j-1}. \quad (4.9)$$

The matrices $M_{i,j} \in \mathbf{R}^{p \times m}$ are called the *Markov parameters* of $T(s, z)$.

Proposition 4.4. $\Sigma = (A_1, A_2, B, C)$ is a realization of $T(s, z)$ if and only if, for any $i, j \in \mathbb{N}$,

$$M_{i,j} = CA_1^i A_2^j B. \quad (4.10)$$

Proof. By (4.7) we have $T(s, z) = C(sI - A_1)^{-1}(zI - A_2)^{-1}B = C\left(\sum_{i=0}^{\infty} A_1^i s^{-i-1}\right)\left(\sum_{j=0}^{\infty} A_2^j z^{-j-1}\right)B = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} CA_1^i A_2^j s^{-i-1} z^{-j-1}$. Since (4.9) holds and two equal Laurent series have equal corresponding coefficients, (4.10) is true.

The following theorem establishes the connection between the concepts of reachability, observability and minimality (the proof is omitted because of lack of space).

Theorem 4.5. A system $\Sigma = (A_1, A_2, B, C)$ is a minimal realization of some $T(s, z) \in SPS(p, m)$ if and only if Σ is completely reachable and completely observable.

5. Minimal realizations

Let us consider a matrix $T(s, z) \in SPS(p, m)$ and its Markov parameters $M_{i,j}, i, j \geq 0$ given by (4.9). We associate to $T(s, z)$ the block Hankel matrices

$$H_k^j = \begin{bmatrix} M_{0,j} & M_{1,j} & \dots & M_{k-1,j} \\ M_{1,j} & M_{2,j} & \dots & M_{k,j} \\ \vdots & \vdots & \ddots & \vdots \\ M_{k-1,j} & M_{k,j} & \dots & M_{2k-2,j} \end{bmatrix}, H_{k,l} = \begin{bmatrix} H_k^0 & H_k^1 & \dots & H_k^{l-1} \\ H_k^1 & H_k^2 & \dots & H_k^l \\ \vdots & \vdots & \ddots & \vdots \\ H_k^{l-1} & H_k^l & \dots & H_k^{2l-2} \end{bmatrix}. \quad (5.1)$$

Proposition 5.1. For any realization Σ of $T(s, z)$ and any $k, l \geq 1$, $\text{rank } H_{k,l} \leq \dim \Sigma$.

Proof. For some realization Σ we denote by $C_{k,l}$ and $O_{k,l}$ the matrices with similar structures as C_Σ (3.3) and O_Σ (3.5) but with kl blocks. Then $O_{k,l}C_{k,l} = H_{k,l}$ and by the second Sylvester Inequality we have for any $k, l \geq 1$, $\text{rank } H_{k,l} \leq \min(\text{rank } O_{k,l}, \text{rank } C_{k,l}) \leq \dim \Sigma$.

Now let $p_1(s)p_2(z)$ be the separable least common denominator of entries of $T(s, z)$, where $p_1(s) = s^q + \sum_{i=0}^{q-1} \alpha_i s^i$, $p_2(z) = z^r + \sum_{j=0}^{r-1} \beta_j z^j$. Let us consider the shift operators $\tilde{\sigma}_1$ (first level), σ_1 and σ_2 (second level) applied to the two level block Hankel matrix $H_{q,r}$, defined by

$$\begin{aligned}\tilde{\sigma}_1 H_q^j &= \begin{bmatrix} M_{1,j} & M_{2,j} & \dots & M_{q,j} \\ M_{2,j} & M_{3,j} & \dots & M_{q+1,j} \\ \vdots & \vdots & \ddots & \vdots \\ M_{q,j} & M_{q+1,j} & \dots & M_{2q-1,j} \end{bmatrix} \\ \sigma_1 H_{q,r} &= \begin{bmatrix} \tilde{\sigma}_1 H_q^0 & \tilde{\sigma}_1 H_q^1 & \dots & \tilde{\sigma}_1 H_q^{r-1} \\ \tilde{\sigma}_1 H_q^1 & \tilde{\sigma}_1 H_q^2 & \dots & \tilde{\sigma}_1 H_q^r \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\sigma}_1 H_q^{r-1} & \tilde{\sigma}_1 H_q^r & \dots & \tilde{\sigma}_1 H_q^{2r-2} \end{bmatrix} \\ \sigma_2 H_{q,r} &= \begin{bmatrix} H_q^1 & H_q^2 & \dots & H_q^r \\ H_q^2 & H_q^3 & \dots & H_q^{r+1} \\ \vdots & \vdots & \ddots & \vdots \\ H_q^r & H_q^{r+1} & \dots & H_q^{2r-1} \end{bmatrix}. \end{aligned} \quad (5.2)$$

Let J_k be the companion cell associated to the polynomial p_k , $k = 1, 2$; we introduce the matrices

$$\begin{aligned}F_1 &= I_r \otimes J_1 \otimes I_p, & F_2 &= J_2 \otimes I_q \otimes I_p, \\ \tilde{F}_1 &= I_r \otimes J_1^T \otimes I_m, & \tilde{F}_2 &= J_2^T \otimes I_q \otimes I_m\end{aligned} \quad (5.3)$$

where \otimes denotes the Kronecker product.

Since $F_1 F_2 = (I_r J_2) \otimes (J_1 I_q) \otimes (I_p I_p) = F_2 F_1$, F_1, F_2 and similarly \tilde{F}_1, \tilde{F}_2 are commutative matrices.

Proposition 5.2.

$$\sigma_k H_{q,r} = F_k H_{k,r} = H_{q,r} \tilde{F}_k, k = 1, 2. \quad (5.4)$$

Proof. Since $p_1(s)T(s, z)$ and $p_2(z)T(s, z)$ are polynomial matrices with respect to s and z respectively, if we replace $T(s, z)$ by its Laurent series (4.9) and we equalize with zero the coefficients of the negative powers of s and z we get two recurrence relations:

$$M_{l+q,j} = - \sum_{i=0}^{q-1} \alpha_i M_{l+i,j}, \forall l, j \geq 0; M_{i,l+r} = - \sum_{j=0}^{r-1} \beta_j M_{i,l+j}, \forall l, i \geq 0.$$

Equalities (5.4) result by a long but direct calculus, based on these relations.

By induction we get

Corollary 5.3.

$$\sigma_k^h H_{q,r} = F_k^h H_{k,r} = H_{q,r} \tilde{F}_k^h, k = 1, 2, \forall h \in \mathbf{N}^*. \quad (5.5)$$

Let us denote by 0_k^h the null matrix with k rows and h columns and by E_k^h the $k \times h$ matrix

$$E_k^h = \begin{cases} \begin{bmatrix} I_k & 0_k^{h-k} \end{bmatrix} & \text{if } h > k \\ I_k & \text{if } h = k \\ \begin{bmatrix} I_k \\ 0_{k-h}^h \end{bmatrix} & \text{if } h < k. \end{cases}$$

Algorithm 5.4. (of minimal realization)

Stage I. Expand $T(s, z)$ in Laurent series (4.9) and determine the Markov parameters $M_{i,j}$.

Stage II. Determine the degrees q and r of $p_1(s)$ and $p_2(z)$ where $p_1(s)p_2(z)$ is the l.c.d. of elements of $T(s, z)$.

Stage III. Write the two level block Hankel matrix $H_{q,r}$ and the matrix

$$K = \begin{bmatrix} I_{\bar{p}} & H_{q,r} \\ 0 & I_{\bar{m}} \end{bmatrix} \text{ where } \bar{p} = pqr \text{ and } \bar{m} = mqr.$$

Stage IV. By applying elementary row operations on the first \bar{p} rows of K and elementary column operations on the last \bar{m} columns of K , transform K into

$$\bar{K} = \begin{bmatrix} P & \bar{H} \\ 0 & M \end{bmatrix} \quad \text{where} \quad \bar{H} = \begin{bmatrix} I_n & 0_n^{\bar{m}-n} \\ 0_{\bar{p}-n}^n & 0_{\bar{p}-n}^{\bar{m}-n} \end{bmatrix}. \quad (5.6)$$

Stage V. Calculate the minimal realization $\Sigma = (A_1, A_2, B, C)$ by the formulae

$$\begin{aligned} A_k &= E_n^{\bar{p}} P [\sigma_k H_{q,r}] M E_{\bar{m}}^n, \quad k=1, 2; \\ B &= E_n^{\bar{p}} P H_{q,r} E_{\bar{m}}^n; \quad C = E_p^{\bar{p}} H_{q,r} M E_{\bar{m}}^n. \end{aligned} \quad (5.7)$$

Proof. By elementary operations considerations it results that P and M are nonsingular since they are products of elementary matrices and that $P H_{q,r} M = \bar{H} = E_{\bar{p}}^n E_{\bar{m}}^{\bar{m}}$. Moreover, the matrix $Q = M E_{\bar{m}}^n E_{\bar{p}}^{\bar{p}} P$ is the pseudoinverse of $H_{q,r}$, i.e.

$$H_{q,r} Q H_{q,r} = H_{q,r}. \quad (5.8)$$

Firstly let us show that A_1 and A_2 are commutative matrices. We have by (5.4) and (5.8) $A_1 A_2 = (E_n^{\bar{p}} P [\sigma_1 H_{q,r}] M E_{\bar{m}}^n)(E_p^{\bar{p}} H_{q,r} M E_{\bar{m}}^n) =$

$E_n^{\bar{p}}PF_1H_{q,r}QH_{q,r}\tilde{F}_2ME_m^n = E_n^{\bar{p}}PF_1H_{q,r}\tilde{F}_2ME_m^n = E_n^{\bar{p}}PF_1F_2H_{q,r}ME_m^n$ and A_2A_1 has the same expression since F_1 and F_2 are commutative matrices.

In a similar way we can prove by induction that

$$A_1^i A_2^j = E_n^{\bar{p}}PF_1^i F_2^j H_{q,r}ME_m^n \quad \forall i, j \geq 0. \quad (5.9)$$

Now we can prove that $M_{i,j} = CA_1^i A_2^j B \quad \forall i, j \geq 0$. Indeed, by (5.7), (5.8), (5.9) and (5.5) we obtain $CB = (E_p^{\bar{p}}H_{q,r}ME_m^n)(E_n^{\bar{p}}PH_{q,r}E_m^m) = E_p^{\bar{p}}H_{q,r}QH_{q,r}E_m^m = E_p^{\bar{p}}H_{q,r}E_m^m = M_{0,0}$; $CA_1^i A_2^j B = (E_p^{\bar{p}}H_{q,r}ME_m^n)(E_n^{\bar{p}}PF_1^i F_2^j H_{q,r}ME_m^n) = (E_p^{\bar{p}}PH_{q,r}E_m^m) = (E_p^{\bar{p}}H_{q,r}MQF_1^i F_2^j H_{q,r}QH_{q,r}E_m^m) = (E_p^{\bar{p}}H_{q,r}MQH_{q,r}\tilde{F}_1^i \tilde{F}_2^j E_m^m) = (E_p^{\bar{p}}\sigma_1^i \sigma_2^j H_{q,r}E_m^m) = M_{i,j} \quad \forall i, j \geq 0$. By Proposition 4.4 Σ (5.7) is a realization of $T(s, z)$.

Since P and M are nonsingular matrices and $PH_{q,r}M = \bar{H}$ it results that $\text{rank}H_{q,r} = \text{rank}\bar{H} = n = \dim\Sigma$. By Proposition 5.1 we get $\dim\Sigma = \text{rank}H_{q,r} \leq \dim\tilde{\Sigma}$ for any realization Σ of $T(s, z)$, hence the realization Σ (5.7) is minimal.

References

- [1] B. De Schutter, "Minimal state-space realization in linear system theory: an overview", *Journal of Computational and Applied Mathematics*, **121** (2000), 331-354.
- [2] G.E. Antoniou, P.N. Paraskevopoulos, and S.J. Varoufakis, "Minimal state-space realization of factorable 2-D transfer functions", *IEEE Trans. Circuits Systems*, **35**, 8 (1988), 1055-1058.
- [3] S.H. Mentzelopoulou and N. J. Theodorou, "n-Dimensional Minimal State-Space Realization", *IEEE Trans. Circuits Systems*, **38**, 3 (1991), 340-343.
- [4] S.A. Miri and J.D. Aplevich, "Modeling and Realization of n-Dimensional Linear Discrete Systems", *Multidimensional Systems and Signal Processing*, **9** (1998), 241-253.
- [5] B. L. Ho and R.E. Kalman, "Effective construction of linear, state-variable models from input/output functions", *Regelungstechnik*, **14**, 12 (1966), 545-548.
- [6] Z. Szabó, P.S.C. Heuberger, J. Bokor and P.M.J. Van den Hof, "Extended Ho-Kalman algorithm for systems represented in generalized orthonormal bases", *Automatica*, **36** (2000), 1809-1818.
- [7] V. Prepelita, "Systèmes linéaires à N indices". *C.R. Acad. Sc. Paris*, 279 (1974), 387-390.
- [8] S. Attasi, "Introduction d'une classe de systèmes linéaires récurrents à deux indices". *C.R. Acad. Sc. Paris*, 277 (1973), 1135.
- [9] V. Prepelita, "Linear hybrid systems", *Bull. Math. Soc. Sci. Math. de Roumanie*, **23** (71), 4 (1979), 391-403.
- [10] R.P. Roesser, "A Discrete State-Space Model for Linear Image Processing", *IEEE Trans. Aut. Control*, **AC-20** (1975), 1-10.

- [11] E. Fornasini and G. Marchesini, "State Space Realization Theory of Two-Dimensional Filters", *IEEE Trans. Aut. Control*, **AC-21** (1976), 484-492.
- [12] T. Kaczorek, "Controllability and minimum energy control of 2D continuous-discrete linear systems", *Appl. Math. and Comp. Sci.*, **5**, 1 (1995), 5-21.
- [13] T. Kaczorek, "Singular Two-Dimensional Continuous-Discrete Linear Systems", *Dynamics of Continuous, Discrete and Impulsive Systems*, **2** (1996), 193-204.
- [14] T. Kaczorek, *Positive 1D and 2D Systems*, Springer, London Berlin Heidelberg, 2002.
- [15] K. Galkowski, E. Rogers and D.H. Owens, "New 2D models and a transition matrix for discrete linear repetitive processes", *Int.J. Control.*, **72**, 15 (1999), 1365-1380.
- [16] J. Kurek and M.B. Zaremba, "Iterative learning control synthesis on 2D system theory", *IEEE Trans. Aut. Control*, **AC-38**, 1 (1993), 121-125.
- [17] P. Picard, J. F. Lafay and V. Kucera, "Model Matching for Linear Systems with Delays and 2D Systems", *Automatica*, **34**, 2 (1998), 183-191.
- [18] V. Prepelită, "Generalized Dynamical systems", *Scientific Bulletin, University Politehnica of Bucharest*, **53**, 3-4 (1991), 257-268.
- [19] V. Prepelită and C. Drăgușin, "Linear Boundary Value Systems in the Space of Regulated Functions", *Qualitative Problems for Differential Equations and Control Theory*, C. Corduneanu (Ed.), World Scientific Publishing Co., (1995), 185-196.
- [20] V. Prepelită, "Calculus of the fundamental matrix for generalized systems of differential equations", *Annales Sci. Math. Québec*, **23**, 1 (1999), 87-96.
- [21] V. Prepelită, "Criteria of reachability for 2D continuous-discrete systems", To appear in *Rev. Roumaine Math. Pures Appl.*
- [22] V. Prepelită and Monica Pârvan, "Observability of 2D continuous-discrete separable systems", To appear in *Rev. Roumaine Math. Pures Appl.*

NUMERICAL METHODS FOR NASH EQUILIBRIA IN MULTIOBJECTIVE CONTROL OF PARTIAL DIFFERENTIAL EQUATIONS*

Angel Manuel Ramos

Departamento de Matemática Aplicada

Universidad Complutense de Madrid

Angel_Ramos@mat.ucm.es

Abstract This paper is concerned with the numerical solution of multiobjective control problems associated with linear (resp., nonlinear) partial differential equations. More precisely, for such problems, we look for Nash equilibria, which are solutions to noncooperative games. First, we study the continuous case. Then, to compute the solution of the problem, we combine finite-difference methods for the time discretization, finite-element methods for the space discretization, and conjugate gradient algorithms (resp., a suitable algorithm) for the iterative solution of the discrete control problems. Finally, we apply the above methodology to the solution of several tests problems.

Keywords: Partial differential equations, Heat equation, Burgers equation, optimal control, pointwise control, Nash equilibria, adjoint systems, conjugate gradient methods, multiobjective optimization, quasi-Newton algorithms.

1. Introduction

In this paper we present some methods for the numerical computation of the solutions of some multiobjective control problems associated with partial differential equations. The details about the results and algorithms showed here can be seen in [8], [9].

In a classical *single-objective* control problem for a system modelled by a Differential Equation, there is an output control v , acting on the

*Partial funding provided by the Spanish 'Plan Nacional de I+D+I (2000-2003) MCYT' through the AGL2000-1440-C02-01 project.

equation and trying to achieve a pre-determined goal, usually consisting of minimizing a functional $J(\cdot)$.

In a *multiobjective* control problem there are more than one goal and, possibly, more than one control acting on the equation. Now, in contrast with the single-objective case, there are several strategies in order to choose the controls, depending of the character of the problem. These strategies can be cooperative (when the controls cooperate between them in order to achieve the goals), non-cooperative, hierarchical, etc..

Nash equilibria define a *noncooperative multiple objective optimization strategy* first proposed by Nash [6]. Since it originated in *game theory* and *economics*, the notion of *player* is often used. For an optimization problem with G objectives (or functionals J_i to minimize), a *Nash strategy* consists in having G *players* (or controls v_i), each optimizing his own criterion. However, each player has to optimize his criterion given that all the other criteria are fixed by the rest of the *players*. When no *player* can further improve his criterion, it means that the system has reached a *Nash Equilibrium* state.

Of course there are other strategies for *multiobjective optimization*, such as the *Pareto* (cooperative) strategy [7] and the *Stackelberg* (hierarchical) strategy [10], etc..

Some previous works about these strategies for the control of partial differential equations are the following: In the articles by Lions [3]-[4] the author gives some results about the Pareto and Stackelberg strategies, respectively. In the article by Díaz and Lions [2], the authors prove an approximate controllability result for a system following a *Stackelberg-Nash strategy*. In the article by Bristeau et al. [1], the authors compare *Pareto and Nash strategies* by using *genetic algorithms* to compute numerically the solutions corresponding to these strategies.

2. Formulation of the problems

2.1. A linear case

Let us consider $T > 0$ and $\Omega \subset I\!\!R^d$, $d = 1$ or 2 . We define $Q = \Omega \times (0, T)$ and $\Sigma = \partial\Omega \times (0, T)$. We define the control spaces $\mathcal{U}_1 = L^2(\omega_1 \times (0, T))$ and $\mathcal{U}_2 = L^2(\omega_2 \times (0, T))$, where $\omega_1, \omega_2 \subset \Omega$ and $\omega_1 \cap \omega_2 = \emptyset$. Finally, we consider the functionals J_1 and J_2 given by

$$\begin{aligned} J_i(v_1, v_2) = & \frac{\alpha_i}{2} \|v_i\|_{\mathcal{U}}^2 + \frac{k_i}{2} \|y(v_1, v_2) - y_{d,i}\|_{L^2(\omega_{di} \times (0, T))}^2 \\ & + \frac{l_i}{2} \|y(v_1, v_2; T) - y_{T,i}\|_{L^2(\omega_{Ti})}^2, \quad i = 1, 2, \end{aligned} \quad (1)$$

for every $(v_1, v_2) \in \mathcal{U}_1 \times \mathcal{U}_2$, where $\omega_{di}, \omega_{Ti} \subset \Omega$ ($i = 1, 2$) and function y is defined as the solution of

$$\begin{cases} \frac{\partial y}{\partial t} - \Delta y = f + v_1 \chi_{\omega_1} + v_2 \chi_{\omega_2} & \text{in } Q, \\ y(x, 0) = y_0(x) & \text{in } \Omega, \\ y = g & \text{on } \Sigma, \end{cases} \quad (2)$$

with $f, g, y_0, y_{d,i}$ and $y_{T,i}$ being smooth enough functions, $\alpha_i > 0$, $k_i, l_i \geq 0$ and $k_i + l_i > 0$ ($i = 1, 2$).

Remark 2.1 The following is also valid for more than two controls (and functionals), for more general linear operators, for different type of controls such as, for instance, boundary or initial controls and for different type of functionals.

Now, for every $w_2 \in \mathcal{U}_2$ we consider the optimal control problem $(\mathcal{CP}_1(w_2))$: Find $u_1(w_2) \in \mathcal{U}_1$, such that

$$J_1(u_1(w_2), w_2) \leq J_1(v_1, w_2), \quad \forall v_1 \in \mathcal{U}_1;$$

similarly for every $w_1 \in \mathcal{U}_1$ we consider the optimal control problem $(\mathcal{CP}_2(w_1))$: Find $u_2(w_1) \in \mathcal{U}_2$, such that

$$J_2(w_1, u_2(w_1)) \leq J_2(w_1, v_2), \quad \forall v_2 \in \mathcal{U}_2.$$

The (unique) solution $u_1(w_2)$ (respectively $u_2(w_1)$) of $(\mathcal{CP}_1(w_2))$ (respectively $(\mathcal{CP}_2(w_1))$) is characterized by $\frac{\partial J_1}{\partial v_1}(u_1(w_2), w_2) = 0$ (respectively $\frac{\partial J_2}{\partial v_2}(w_1, u_2(w_1)) = 0$).

A Nash equilibrium is a pair $(u_1, u_2) \in \mathcal{U}_1 \times \mathcal{U}_2$ such that $u_1 = u_1(u_2)$ and $u_2 = u_2(u_1)$, i.e. (u_1, u_2) is a solution of the *coupled system*:

$$\begin{cases} \frac{\partial J_1}{\partial v_1}(u_1, u_2) = 0 \\ \frac{\partial J_2}{\partial v_2}(u_1, u_2) = 0. \end{cases} \quad (3)$$

We show that system (3) has a unique solution. Furthermore, we give a numerical method for the solution of this problem and present the results obtained with this method on some examples.

Remark 2.2 A special case is when $\omega_{T1} \cap \omega_{T2} \neq \emptyset$ and/or $\omega_{d1} \cap \omega_{d2} \neq \emptyset$. This case is a *competition-wise* problem, with each control (or *player*) trying to reach (possibly) different goals over a common *domain*. In some sense this is the case where the behavior of the solution y associated to the equilibrium (u_1, u_2) is most difficult to forecast.

It is obvious that the mapping

$$\begin{aligned} \left(\frac{\partial J_1}{\partial v_1}, \frac{\partial J_2}{\partial v_2} \right) : (v_1, v_2) \in \mathcal{U}_1 \times \mathcal{U}_2 &\longrightarrow \\ \longrightarrow \left(\frac{\partial J_1}{\partial v_1}(v_1, v_2), \frac{\partial J_2}{\partial v_2}(v_1, v_2) \right) &\in \mathcal{U}_1 \times \mathcal{U}_2 \end{aligned} \quad (4)$$

is an affine mapping of $V := \mathcal{U}_1 \times \mathcal{U}_2$. Therefore, there exist a linear continuous mapping $\mathcal{A} \in \mathcal{L}(V, V)$ and a vector $b \in V$ such that

$$\left(\frac{\partial J_1}{\partial v_1}(v_1, v_2), \frac{\partial J_2}{\partial v_2}(v_1, v_2) \right) = \mathcal{A}(v_1, v_2) - b.$$

Let us identify mapping \mathcal{A} : For every $(v_1, v_2) \in V$, the linear part of the affine mapping in relation (4) is defined by

$$\mathcal{A}(v_1, v_2) = (\alpha_1 v_1 + p_1 \chi_{\omega_1}, \alpha_2 v_2 + p_2 \chi_{\omega_2}),$$

where p_i , $i = 1, 2$, is the solution of

$$\begin{cases} -\frac{\partial p_i}{\partial t} - \Delta p_i = k_i y \chi_{\omega_{di}} & \text{in } Q, \\ p_i(x, T) = l_i y(T) \chi_{\omega_{Ti}} & \text{in } \Omega, \\ p_i = 0 & \text{on } \Sigma, \end{cases}$$

and y is the solution of (2) with $f \equiv 0$, $y_0 \equiv 0$ and $g \equiv 0$.

Proposition 2.1 Mapping \mathcal{A} is linear, continuous, symmetric and strongly positive.

Let us identify b : The constant part of the affine mapping (4) is the function $b \in V$ defined by $b = (p_1 \chi_{\omega_1}, p_2 \chi_{\omega_2})$, where p_i , $i = 1, 2$, is the solution of

$$\begin{cases} -\frac{\partial p_i}{\partial t} - \Delta p_i = k_i (Y - y_{d,i}) \chi_{\omega_{di}} & \text{in } Q, \\ p_i(x, T) = l_i (Y(T) - y_{T,i}) \chi_{\omega_{Ti}} & \text{in } \Omega, \\ p_i = 0 & \text{on } \Sigma, \end{cases}$$

and Y is the solution of (2) with $v_1 = 0$ and $v_2 = 0$.

Now, if we define $a(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$ by

$$a(v, w) = (\mathcal{A}(v), w)_V \quad \forall v, w \in V,$$

and $L : V \rightarrow \mathbb{R}$ by

$$L(v) = (b, v)_V, \quad \forall v \in V,$$

Proposition 2.1 proves that mapping $a(\cdot, \cdot)$ is bilinear continuous, symmetric and V -elliptic; mapping L is (obviously) linear and continuous. Thus, system (3) has a unique solution, which can be computed by the following *conjugate gradient* algorithm:

Step 1. (u_1^0, u_2^0) is given in V .

Step 2.a. y^0 is the solution of (2) with $v_1 = u_1^0$ and $v_2 = u_2^0$.

$$\text{Step 2.b. For } i = 1, 2, \begin{cases} -\frac{\partial p_i^0}{\partial t} - \Delta p_i^0 = k_i(y^0 - y_{i,d})\chi_{\omega_{di}} & \text{in } Q, \\ p_i^0(x, T) = l_i((y^0(T) - y_{i,T})\chi_{\omega_{Ti}}) & \text{in } \Omega, \\ p_i^0 = 0 & \text{on } \Sigma. \end{cases}$$

Step 2.c. $(g_1^0, g_2^0) = (\alpha_1 u_1^0 + p_1^0 \chi_{\omega_1}, \alpha_2 u_2^0 + p_2^0 \chi_{\omega_2}) \in V$.

Step 3. $(w_1^0, w_2^0) = (g_1^0, g_2^0) \in V$.

For $k \geq 0$, assuming that (u_1^k, u_2^k) , (g_1^k, g_2^k) , (w_1^k, w_2^k) are known, we compute (u_1^{k+1}, u_2^{k+1}) , (g_1^{k+1}, g_2^{k+1}) and (if necessary) (w_1^{k+1}, w_2^{k+1}) as follows:

Step 4.a. \bar{y}^k is the solution of (2) with $f \equiv 0, y_0 \equiv 0, g \equiv 0, v_1 = w_1^k$ and $v_2 = w_2^k$.

$$\text{Step 4.b. For } i = 1, 2, \begin{cases} -\frac{\partial \bar{p}_i^k}{\partial t} - \Delta \bar{p}_i^k = k_i \bar{y}^k \chi_{\omega_{di}} & \text{in } Q, \\ \bar{p}_i^k(x, T) = l_i \bar{y}^k(T) \chi_{\omega_{Ti}} & \text{in } \Omega, \\ \bar{p}_i^k = 0 & \text{on } \Sigma. \end{cases}$$

Step 4.c. $(\bar{g}_1^k, \bar{g}_2^k) = (\alpha_1 w_1^k + \bar{p}_1^k \chi_{\omega_1}, \alpha_2 w_2^k + \bar{p}_2^k \chi_{\omega_2})$.

$$\text{Step 4.d. } \rho_k = \frac{\| (g_1^k, g_2^k) \|_V^2}{\int_{\omega_1 \times (0, T)} \bar{g}_1^k w_1^k dx dt + \int_{\omega_2 \times (0, T)} \bar{g}_2^k w_2^k dx dt}.$$

Step 5. $(u_1^{k+1}, u_2^{k+1}) = (u_1^k, u_2^k) - \rho_k (w_1^k, w_2^k)$.

Step 6. $(g_1^{k+1}, g_2^{k+1}) = (g_1^k, g_2^k) - \rho_k (\bar{g}_1^k, \bar{g}_2^k)$.

If $\frac{\| (g_1^{k+1}, g_2^{k+1}) \|_V^2}{\| (g_1^k, g_2^k) \|_V^2} \leq \varepsilon$, then take $(u_1, u_2) = (u_1^{k+1}, u_2^{k+1})$; else:

$$\text{Step 7. } \gamma_k = \frac{\| (g_1^{k+1}, g_2^{k+1}) \|_V^2}{\| (g_1^k, g_2^k) \|_V^2}.$$

Step 8. $(w_1^{k+1}, w_2^{k+1}) = (g_1^{k+1}, g_2^{k+1}) + \gamma_k (w_1^k, w_2^k)$.

Step 9. Do $k = k + 1$, and go to Step 4.a.

2.2. A nonlinear case

We shall consider the Burgers equation with pointwise controls. All the results to follow are also valid for more than two control points but for simplicity we shall consider the case of only two control points a_1 and a_2 . Let $Q = (0, 1) \times (0, T)$. The state equation is

$$\begin{cases} y_t - \nu y_{xx} + yy_x = f + v_1\delta(x - a_1) + v_2\delta(x - a_2) & \text{in } Q, \\ y_x(0, t) = 0, \quad y(1, t) = 0 & \text{in } (0, T), \\ y(0) = y_0 & \text{in } (0, 1). \end{cases} \quad (5)$$

Let us consider $\omega_{di}, \omega_{Ti} \subset (0, 1)$ ($i = 1, 2$) and the target functions $y_{di} \in L^2(\omega_d \times (0, T))$ and $y_{Ti} \in L^2(\omega_T)$ ($i = 1, 2$). We take as the control space $\mathcal{U}_1 = \mathcal{U}_2 = \mathcal{U} = L^2(0, T)$.

The goal of each control v_i ($i = 1, 2$) is to drive the solution y close to y_{di} in $\omega_{di} \times (0, T)$ and $y(T)$ close to y_{Ti} in ω_{Ti} at a minimal cost for the control v_i . To do this, we define again two *cost functions* $J_i(v_1, v_2)$ as in (1).

For every $w_1 \in \mathcal{U}_1$ and $w_2 \in \mathcal{U}_2$ we consider the optimal control problems $(\mathcal{CP}_1(w_2))$ and $(\mathcal{CP}_2(w_1))$ as before. A Nash equilibrium is a pair $(u_1, u_2) \in \mathcal{U}_1 \times \mathcal{U}_2$ such that $u_1 = u_1(u_2)$ and $u_2 = u_2(u_1)$.

The algorithm we propose is the following:

Step 1. (u_1^0, u_2^0) is given in $\mathcal{U}_1 \times \mathcal{U}_2$.

Step 2. We get u_1^1 as the solution of $(\mathcal{CP}_1(u_2^0))$.

Step 3. We get u_2^1 as the solution of $(\mathcal{CP}_2(u_1^0))$.

Then, for $k \geq 1$, assuming that $(u_1^k, u_2^k) \in \mathcal{U}_1 \times \mathcal{U}_2$ is known, we compute (u_1^{k+1}, u_2^{k+1}) as follows:

Step 4. If $u_2^k = u_2^{k-1}$ then $u_1^{k+1} = u_1^k$;

else get u_1^{k+1} as the solution of $(\mathcal{CP}_1(u_2^k))$.

Step 5. If $u_1^k = u_1^{k-1}$ then $u_2^{k+1} = u_2^k$;

else get u_2^{k+1} as the solution of $(\mathcal{CP}_2(u_1^k))$.

Step 6. If $u_1^{k+1} = u_1^k$ and $u_2^{k+1} = u_2^k$ then take $(u_1, u_2) = (u_1^{k+1}, u_2^{k+1})$;

else do $k = k + 1$ and go to Step 4.

Most of the descent methods for the numerical solution of $(\mathcal{CP}_i(u_j^k))$ will require the solution of the corresponding gradient, which we can be easily determined by a suitable adjoint system as in the previous linear case.

The following Remark is valid for both linear and nonlinear cases.

Remark 2.3 If $y_{d,1} = y_{d,2} = y_d$, $y_{T,1} = y_{T,2} = y_T$, $\alpha_1 = \alpha_2 = \alpha$, $k_1 = k_2 = k$ and $l_1 = l_2 = l$, then the Nash Equilibria problem (3) is equivalent to the classical control problem (\mathcal{CP}) : Find $(u_1, u_2) \in \mathcal{U}_1 \times \mathcal{U}_2$, such that

$$J(u_1, u_2) \leq J(v_1, v_2), \quad \forall (v_1, v_2) \in \mathcal{U}_1 \times \mathcal{U}_2,$$

where

$$J(v_1, v_2) = \frac{\alpha}{2} \| v \|_{\mathcal{U}}^2 + \frac{k}{2} \| y - y_d \|_{L^2(\omega_d \times (0, T))}^2 + \frac{l}{2} \| y(T) - y_T \|_{L^2(\omega_T)}^2.$$

3. Time discretizations

For simplicity, we consider from now on the special *competition-wise* control problem (see Remark 2.2) given by the case where $k_1 = k_2 = k$, $l_1 = l_2 = l$, $\omega_{d1} = \omega_{d2} = \omega_d$ and $\omega_{T1} = \omega_{T2} = \omega_T$.

3.1. Linear case

We point out that, for the special case specified above, the mapping \mathcal{A} defined in Section 2.1 is $\mathcal{A}(v_1, v_2) = (\alpha_1 v_1 + p \chi_{\omega_1}, \alpha_2 v_2 + p \chi_{\omega_2})$, with $p = p_1 = p_2$ (since p_1 and p_2 are solution of the same equation). Further, the functions \bar{p}_1^k and \bar{p}_2^k defined in the Step 4.b of the *Conjugate Gradient* algorithm are solution of the same equation and therefore $\bar{p}_1^k = \bar{p}_2^k = \bar{p}^k$.

We consider the time discretization step Δt , defined by $\Delta t = T/N$, where N is a positive integer. Then, if we denote $n\Delta t$ by t^n , we have $0 < t^1 < t^2 < \dots < t^N = T$. For simplicity, we assume that $f, g, y_{d,1}$ and $y_{d,2}$ are continuous functions, at least with respect to the time variable (if not we can always use continuous approximations of these functions). Now, we approximate \mathcal{U}_i by $\mathcal{U}_i^{\Delta t} = (L^2(\omega_i))^N$, $i = 1, 2$. Then, for every $w_2 \in \mathcal{U}_2^{\Delta t}$ we approximate problem $(\mathcal{CP}_1(w_2))$ by the following minimization problem $(\mathcal{CP}_1(w_2))^{\Delta t}$: Find $u_1^{\Delta t}(w_2) \in \mathcal{U}_1^{\Delta t}$, such that

$$J_1^{\Delta t}(u_1^{\Delta t}(w_2), w_2) \leq J_1^{\Delta t}(v_1, w_2), \quad \forall v_1 \in \mathcal{U}_1^{\Delta t},$$

with

$$\begin{aligned} J_1^{\Delta t}(v_1, v_2) = & \alpha_1 \frac{\Delta t}{2} \sum_{n=1}^N \int_{\omega_1} |v_1^n|^2 dx \\ & + k \frac{\Delta t}{2} \sum_{n=1}^N \int_{\omega_d} |y^n - y_{d,1}(t^n)|^2 dx + \frac{l}{2} \int_{\omega_T} |y^N - y_{T,1}|^2 dx, \end{aligned}$$

where $\{y^n\}_{n=1}^N$ is defined by the solution of the following semi-discrete parabolic problem:

$$y^0 = y_0, \tag{6}$$

and for $n = 1, \dots, N$,

$$\begin{cases} \frac{y^n - y^{n-1}}{\Delta t} - \Delta y^n = f(t^n) + v_1^n \chi_{\omega_1} + v_2^n \chi_{\omega_2} & \text{in } \Omega, \\ y^n = g(t^n) & \text{in } \partial\Omega. \end{cases} \tag{7}$$

Similarly, for every $w_1 \in \mathcal{U}_1^{\Delta t}$, we approximate problem $(\mathcal{CP}_2(w_1))$ by a minimization problem $(\mathcal{CP}_2(w_1))^{\Delta t}$. Now, it can be proved that

$$\frac{\partial}{\partial v_i} J_i^{\Delta t}(v_1, v_2) = \{\alpha_i v_i^n + p_i^n \chi_{\omega_i}\}_{n=1}^N, \tag{8}$$

for $i = 1, 2$, where $p_i^{N+1} = l(y^N(v_1, v_2) - y_{T,i})\chi_{\omega_T}$, and for $n = N, \dots, 1$,

$$\begin{cases} \frac{p_i^n - p_i^{n+1}}{\Delta t} - \Delta p_i^n = k(y^n(v_1, v_2) - y_{d,i}(t^n))\chi_{\omega_d} & \text{in } \Omega, \\ p_i^n = 0 & \text{on } \partial\Omega. \end{cases}$$

3.2. Nonlinear case

We approximate \mathcal{U} by $\mathcal{U}^{\Delta t} = \mathbb{R}^N$ and problem $(\mathcal{CP}_1(w_2))$ by the following finite-dimensional minimization problem $(\mathcal{CP}_1(w_2))^{\Delta t}$: Find $u_1^{\Delta t}(w_2) = \{u_1^n\}_{n=1}^{N+1} \in \mathcal{U}^{\Delta t}$, such that

$$J_1^{\Delta t}(u_1^{\Delta t}, w_2) \leq J_1^{\Delta t}(v_1, w_2), \quad \forall v_1 = \{v_1^n\}_{n=1}^{N+1} \in \mathcal{U}^{\Delta t},$$

$$\begin{aligned} J_1^{\Delta t}(v_1, v_2) = & \alpha_1 \frac{\Delta t}{2} \sum_{n=1}^N |v_1^n|^2 + \frac{k\Delta t}{2} \sum_{n=1}^N \|y^n - y_{d,1}(n\Delta t)\|_{L^2(\omega_{d1})}^2 \\ & + \frac{l}{2} \left((1-\theta) \|y^{N-1} - y_{T,1}\|_{L^2(\omega_{T1})}^2 + \theta \|y^N - y_{T,1}\|_{L^2(\omega_{T1})}^2 \right), \end{aligned}$$

where $\theta \in (0, 1]$ and $\{y^n\}_{n=1}^N$ is defined from the solution of the following *second order accurate time discretization* scheme of (5):

$$\begin{cases} y^0 = y_0, \\ \frac{y^1 - y^0}{\Delta t} - \nu \frac{\partial^2}{\partial x^2} \left(\frac{2}{3}y^1 + \frac{1}{3}y^0 \right) + y^0 \frac{\partial y^0}{\partial x} \\ \quad = f^1 + \frac{2}{3} \sum_{m=1}^2 v_m^1 \delta(x - a_m) & \text{in } (0, 1), \\ \frac{\partial y^1}{\partial x}(0) = 0, \quad y^1(1) = 0, \end{cases}$$

and for $n \geq 2$,

$$\begin{cases} \frac{\frac{3}{2}y^n - 2y^{n-1} + \frac{1}{2}y^{n-2}}{\Delta t} - \nu \frac{\partial^2}{\partial x^2} y^n + (2y^{n-1} - y^{n-2}) \frac{\partial}{\partial x} (2y^{n-1} - y^{n-2}) \\ \quad = f^n + \sum_{m=1}^2 v_m^n \delta(x - a_m) & \text{in } (0, 1), \\ \frac{\partial y^n}{\partial x}(0) = 0, \quad y^n(1) = 0. \end{cases}$$

Similarly, we approximate $(\mathcal{CP}_2(w_1))$ by $(\mathcal{CP}_2(w_1))^{\Delta t}$. Again, the corresponding gradients can be computed by suitable adjoint systems.

4. Numerical experiments

In order to carry out numerical experiments we fully discretize the problems by adding a Finite Element Method to the time discretizations.

4.1. Linear case

We consider $\Omega = (0, 1) \times (0, 1)$, $\omega_1 = (0, 0.25) \times (0, 0.25)$, $\omega_2 = (0.75, 1) \times (0, 0.25)$, $\omega_T = \Omega$ and $\omega_d = (0.25, 0.75) \times (0.25, 0.75)$ (see Figure 1). The space discretization step h is defined by $h = 1/(I - 1)$, where I is a positive integer. Then, for every $i, j \in \{1, \dots, I\}$, we take the triangulation \mathcal{T}_h with vertex $x_{i,j} = ((i-1)h, (j-1)h)$ and the triangles as in the typical case showed in Figure 2.

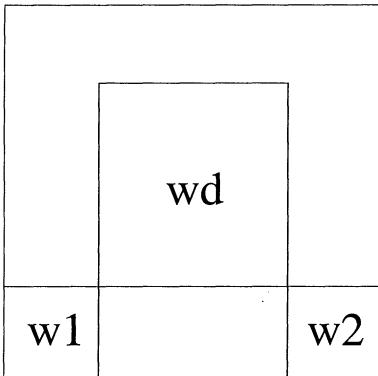


Figure 1. Control and observability domains of the problem.

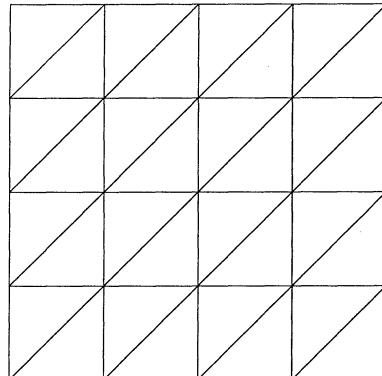


Figure 2. Typical finite element triangulation of Ω .

For the data of the problem we take $f \equiv 1$, $y_0 \equiv 0$ and $g = 0$. In the conjugate gradient algorithm we take the initial guess $(u_1^0, u_2^0) = (0, 0)$ and the stopping criterion $\varepsilon = 10^{-8}$.

We consider the Stabilization Type Test Problem $k = 1$, $l = 0$ with finite horizon time $T = 1.5$, $\Delta t = 1.5/45$ and $h = 1/36$. In order to see how the non-controlled solution behaves, we have visualized in Figure 3 the computed solution of the non-controlled equation at time $t = 1.5$.

We consider the case of Different Goals: $y_{d,1} = 1$, $y_{d,2} = -1$. In Figure 4 we have visualized the graph of the computed solution of the controlled equation with $\alpha_1 = \alpha_2 = 10^{-6}$.

In Figures 5–6 we have visualized the graph of $\|y(t) - 1\|_{L^2(\omega_d)}^2$ and $\|y(t) - (-1)\|_{L^2(\omega_d)}^2$ for different cases. In Table 1 we give some further results about our solution.

Remark 4.1 We point out (see Figures 5–6 and Table 1) that, when the goals are different, the controlled solution can be worse, with respect to both goals, than the uncontrolled solution.

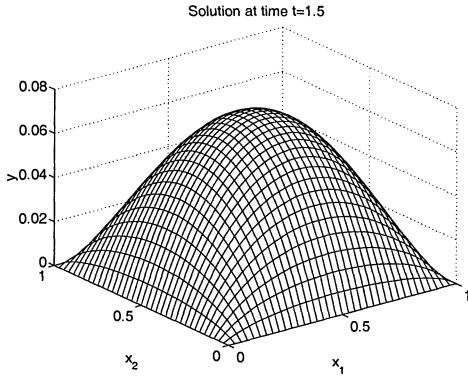


Figure 3. Noncontrolled solution at time $t = 1.5$.

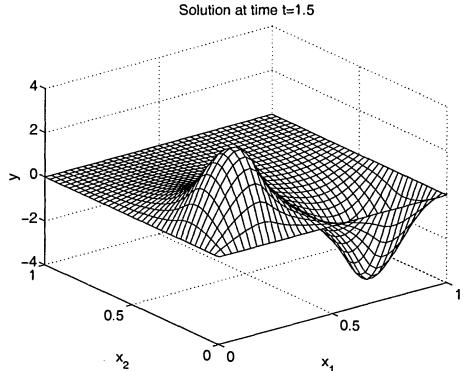


Figure 4. Computed solution of the controlled equation with $\alpha_1 = \alpha_2 = 10^{-6}$ at time $t = 1.5$.

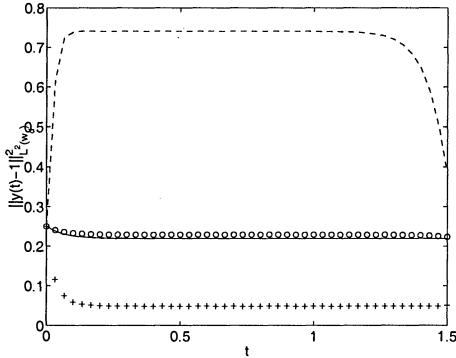


Figure 5. $\|y(t) - 1\|_{L^2(\omega_d)}^2$, y is the computed solution for the following cases: uncontrolled equation (-), $\alpha_1 = \alpha_2 = 10^{-4}$ (oo), $\alpha_1 = \alpha_2 = 10^{-6}$ (- -), $\alpha_1 = 10^{-8}$ and $\alpha_2 = 10^{-2}$ (++).

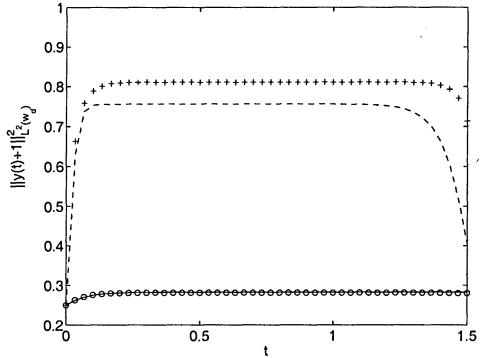


Figure 6. $\|y(t) + 1\|_{L^2(\omega_d)}^2$, y is the computed solution for the following cases: uncontrolled equation (-), $\alpha_1 = \alpha_2 = 10^{-4}$ (oo), $\alpha_1 = \alpha_2 = 10^{-6}$ (- -), $\alpha_1 = 10^{-8}$ and $\alpha_2 = 10^{-2}$ (++).

4.2. Nonlinear case

We consider $T = 1$, $a_1 = 1/5$, $a_2 = 3/5$, $I = 128$, $N = 256$, $\nu = 10^{-2}$,

$$f(x, t) = \begin{cases} 1 & \text{if } (x, t) \in (0, 1/2) \times (0, T), \\ 2(1-x) & \text{if } (x, t) \in [1/2, 1] \times (0, T), \end{cases}$$

$y_0 \equiv 0$ and $\theta = 3/2$. On each minimization problem of the algorithm, we get the sequence u^k ($k = 1, 2, \dots$) by using a quasi-Newton algorithm

Table 1. Computational results for $y_{d,1} = 1$, $y_{d,2} = -1$.

	No control	$\alpha_1 = 10^{-4}$ $\alpha_2 = 10^{-4}$	$\alpha_1 = 10^{-8}$ $\alpha_2 = 10^{-2}$	$\alpha_1 = 10^{-6}$ $\alpha_2 = 10^{-6}$
$\ y(t) - 1\ _{L^2(\omega_d \times (0,1.5))}^2$	0.330592	0.343811	0.0763473	1.07423
$\ y(t) + 1\ _{L^2(\omega_d \times (0,1.5))}^2$	0.422275	0.420292	1.20273	1.09692

à la BFGS (see [5]). We stop iterating after step k if either

$$\begin{aligned} & \left\| \frac{\partial J_h^{\Delta t}}{\partial v}(u^k) \right\|_\infty \leq 10^{-5}, \quad \text{or} \\ & \frac{J_h^{\Delta t}(u^{k-1}) - J_h^{\Delta t}(u^k)}{\max\{|J_h^{\Delta t}(u^{k-1})|, |J_h^{\Delta t}(u^k)|, 1\}} \leq 2 \cdot 10^{-9}. \end{aligned}$$

We consider the Controllability Type Test Problem $\alpha_1 = \alpha_2 = 1$, $k = 0$, $l = 8$. For the case $y_{T1}(x) = \frac{1}{2}(1 - x^3)$, $y_{T2}(x) = 1 - x^3$, Figure 7 shows the uncontrolled state solution $y(T)$ (...), the target functions y_{T1} (- - -), y_{T2} (- . -), and the controlled state solution $y(T)$ (—), when controlling with a Nash strategy. Figure 8 shows the computed controls. In Table 2 we give some further information about several tests.

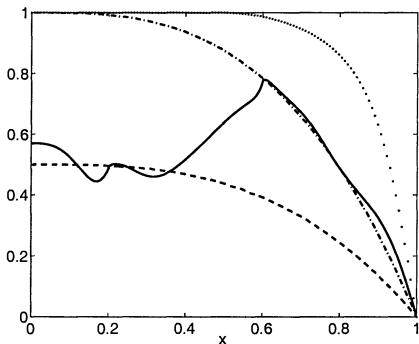


Figure 7. The target functions y_{T1} (- - -), y_{T2} (- . -), the uncontrolled (..) and controlled (—) states, for the Nash strategy, at time T .

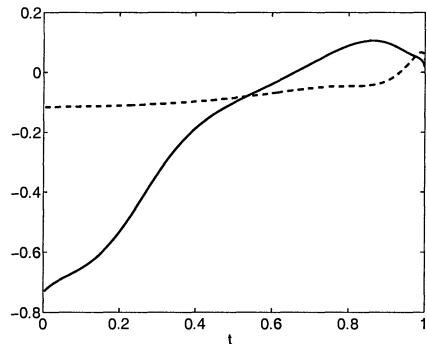


Figure 8. The computed controls u_1 (—) and u_2 (- -) for the Nash strategy.

Acknowledgments

The author wishes to thank Prof. R. Glowinski and Dr. J. Periaux for their encouragement and support.

Table 2. Computational results for the Nash strategy. NQNM= Number of times the Quasi-Newton Method has been used for each functional. NPES= Number of parabolic equations solved for each functional. Test 1: $y_{T1}(x) = y_{T2}(x) = 1 - x^3$. Test 2: $y_{T1}(x) = \frac{1}{2}(1 - x^3)$ and $y_{T2}(x) = 1 - x^3$. Test 3: $y_{T1}(x) = 1 - x^3$ and $y_{T2}(x) = \frac{9}{8}(1 - x^6)$. Test 4: $y_{T1}(x) = \frac{9}{8}(1 - x^6)$ and $y_{T2}(x) = 1 - x^3$.

	Test 1	Test 2	Test 3	Test 4
NQNM J_1 / J_2	19 / 18	5 / 4	6 / 6	55 / 55
NPES J_1 / J_2	286 / 232	188 / 54	78 / 76	1576 / 700
$\ y(0; T) - y_{T1}\ $	0.2522	1.3308	0.2522	0.1001
$\ y_{T1}\ $				
$\ y(u; T) - y_{T1}\ $	0.0241	0.4921	0.2288	0.1702
$\ y_{T1}\ $				
$\ y(0; T) - y_{T2}\ $	0.2522	0.2522	0.1001	0.2522
$\ y_{T2}\ $				
$\ y(u; T) - y_{T2}\ $	0.0241	0.4110	0.1445	0.2395
$\ y_{T2}\ $				
$\ u_1\ $	0.0540	0.3371	0.1334	1.1486
$\ u_2\ $	0.0944	0.0850	0.0849	0.9983

References

- [1] Bristeau, M.O., Glowinski, R., Mantel, B., Periaux, J., and Sefrioui, M., Genetic Algorithms for Electromagnetic Backscattering Multiobjective Optimization. In *Electromagnetic Optimization by Genetic Algorithms*, Edited by Y. Rahmat-Samii and E. Michielssen, John Wiley, New York, pp. 399–434, 1999.
- [2] Díaz, J.I. and Lions, J.L., On the Approximate Controllability of Stackelberg-Nash Strategies. In *Mathematics and Environment*, Lecture Notes, Springer-Verlag, 2001.
- [3] Lions, J.L., Contrôle de Pareto de Systèmes Distribués: Le Cas d'Évolution, *Comptes Rendus de l'Académie des Sciences, Serie I*, 302, 413–417, 1986.
- [4] Lions, J.L. Some Remarks on Stackelberg's Optimization, *Mathematical Models and Methods in Applied Sciences*, 4, 477–487, 1994.
- [5] Liu, D.C., and Nocedal, J., On the Limited Memory BFGS Method for Large-Scale Optimization, *Mathematical Programming*, 45, 503–528, 1989.
- [6] Nash, J.F., Noncooperative Games, *Annals of Mathematics*, 54, 286–295, 1951.
- [7] Pareto, V., *Cours d'Économie Politique*, Rouge, Lausanne, Switzerland, 1896.
- [8] Ramos, A.M., Glowinski, R., and Periaux, J., Nash Equilibria for the Multiobjective Control of Linear Partial Differential Equations, *Journal of Optimization, Theory and Applications*, 112, No. 3, 457–498, 2002.
- [9] Ramos, A.M., Glowinski, R., and Periaux, J., Pointwise Control of the Burgers Equation and Related Nash Equilibrium Problems: Computational Approach, *Journal of Optimization, Theory and Applications*, 112, No. 3, 499–516, 2002.
- [10] von Stackelberg, H., *Marktform und Gleichgewicht*, Springer, Berlin, Germany, 1934.

LOCAL BIFURCATION FOR THE FITZHUGH-NAGUMO SYSTEM

Carmen Rocsoreanu

Dept. of Math., University of Craiova, A.I. Cuza 13, Romania

carmenr@central.ucv.ro

Mihaela Sterpu

Dept. of Math., University of Craiova, A.I. Cuza 13, Romania

mihaelas@central.ucv.ro

Abstract The 2-D FitzHugh-Nagumo (F-N) system depending on three real parameters a , b , and c is considered. It models the electrical potential of the nodal system in the heart. All local bifurcations of equilibria are emphasized in three qualitatively distinct situations concerning the parameter c ($0 < c < 1$, $c = 1$, $c > 1$). We found codimension-one bifurcations (saddle-node, Hopf), codimension two bifurcations (Bogdanov-Takens, Bautin, cusp, double-zero with order two symmetry) and a codimension three bifurcation (degenerated Bogdanov-Takens of order two). In addition, some non-generic codimension two bifurcations generated by the coexistence of two codimension one bifurcations are shown. In our study we used the normal form theory [3], [6] and the center manifold theory [2].

Keywords: bifurcation, FitzHugh-Nagumo, normal form, center manifold.

Introduction

Starting with the Hodgkin-Huxley model [5] for nerve axon which consists of a Cauchy problem for a system of four first order ordinary differential equations (ODE), FitzHugh [4] constructed a two dimensional model, describing physiological phenomena that take place in the sinus node of the heart. This node produces the initiation of the electrical impulse, so it has the role of pace-maker.

The model considered by FitzHugh [4] is the Cauchy problem for the system of ODE's

$$\begin{cases} \frac{dx}{dt} = c \left(x + y + z - \frac{x^3}{3} \right), \\ \frac{dy}{dt} = -\frac{1}{c}(x - a + by), \end{cases} \quad (1)$$

where x is the electrical potential of the cell membrane, y is an auxiliary variable depending on the refractory period, the parameters a and b are related to the number of channels of the cell membrane which are opened to the Na^+ and K^+ ions, z is the injected current and $c \neq 0$ is the relaxation parameter.

For $a = b = z = 0$, the well-known Van der Pol system is obtained.

FitzHugh analyzed this model for some particular values of the parameters a , b , c and variable z , emphasizing the periods of the cardiac cycle.

In this paper we synthetize our results on local bifurcation around equilibria of the F-N model, for all values of the parameters. The system (1) is reduced to the form

$$\begin{cases} \dot{x} = c \left(x + \bar{y} - \frac{x^3}{3} \right), \\ \dot{\bar{y}} = -\frac{1}{c}(x - \bar{a} + b\bar{y}), \end{cases} \quad (2)$$

where $\bar{y} = y + z$, $\bar{a} = a + bz$. We rename the variables x , \bar{y} as x_1 and x_2 , and the parameter \bar{a} as a , so system (2) becomes

$$\begin{cases} \dot{x}_1 = c(x_1 + x_2 - x_1^3/3), \\ \dot{x}_2 = -\frac{1}{c}(x_1 + bx_2 - a), \end{cases} \quad (3)$$

As the system (3) is invariant with respect to the transformations $(x_1, x_2, a) \rightarrow (-x_1, -x_2, -a)$ and $(x_1, x_2, t, c) \rightarrow (x_1, x_2, -t, -c)$, it is sufficient to investigate the case $c > 0$, $a \geq 0$.

Since its derivation, the F-N system was widely investigated. In more particular or general settings it is a subject of many recent articles. The system has been analyzed within a particular parameter region relevant to physiology [4] and for full parameter space with emphasis on the periodic solutions [1]. A complete study on the phase dynamics and bifurcation of this model for $c > 1 + \sqrt{3}$ can be found in [7]. In [10], [12] codimension-one, -two, -three local bifurcations for the entire parameter space are determined.

1. Equilibria and linear stability

Denote by $D = \{ \mu = (b, a, c) \in \mathbf{R}^3, c \neq 0 \}$. For $\mu \in D$, denote by S_μ the set of equilibria of system (3) corresponding to μ . This set is defined by the equations

$$\begin{cases} x_1 + x_2 - \frac{1}{3}x_1^3 = 0, \\ x_1 + bx_2 - a = 0. \end{cases} \quad (4)$$

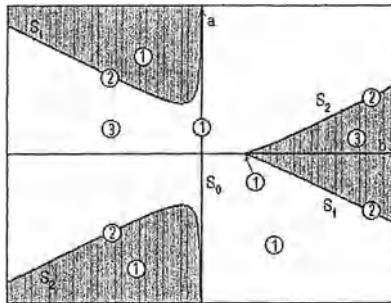


Figure 1. Section in static bifurcation values set with a $c = \text{const.}$ plane.

Let $f : \mathbf{R} \rightarrow \mathbf{R}$ be the function defined by $f(x_1) = x_1^3/3 - x_1$.

Remark 1. The correspondence $(x_1, x_2) \rightarrow (x_1, f(x_1))$ in \mathbf{S}_μ is bijective [7]. Therefore, the static bifurcation diagram of (3) is the set

$$\Sigma = \{(\mu, x_1), \quad \mu \in D, \quad \mathbf{x} = (x_1, x_2) \in \mathbf{S}_\mu\} \subset \mathbf{R}^4.$$

For $b = 0$, the system (4) possesses a unique solution, $x_1 = a$, $x_2 = a^3/3 - a$.

For $b \neq 0$, the abscissa x_1 of an equilibrium satisfies the following equation

$$x_1^3 - 3x_1(1 - \frac{1}{b}) - \frac{3a}{b} = 0. \quad (5)$$

Denote by $\Delta = 4(1 - \frac{1}{b})^3 - \frac{9a^2}{b^2}$ the discriminant of (5). Equation (5) possesses (i) three distinct real solutions $\xi_1 < \xi_2 < \xi_3$ if $\Delta > 0$, (ii) two distinct real solutions (one of them being double) if $\Delta = 0$, (iii) a unique real solution ξ_0 otherwise. The solution ξ_0 is triple for $b = 1$, $a = 0$ and single otherwise.

In the parameter space D , consider the surfaces

$$S_{1,2} : a = \pm \frac{2b}{3}(1 - \frac{1}{b})^{3/2}, b \in (-\infty, 0) \cup [1, \infty) \quad (6)$$

of parameters to which a double solution of equation (5) corresponds, and $S_0 : b = 0$. The static bifurcation diagram for the F-N system (3) consists of $S_1 \cup S_2 \cup S_0$. A section with a $c = \text{const.}$ plane in the static bifurcation diagram is represented in Figure 1., where the number of equilibria corresponding to each strata is shown.

Remark 2. For $\mu \in S_1 \cup S_2$ we have $\xi_1 = -2\sqrt{1 - \frac{1}{b}}$, $\xi_2 = \xi_3 = \sqrt{1 - \frac{1}{b}}$, for $b < 0$, and $\xi_1 = \xi_2 = -\sqrt{1 - \frac{1}{b}}$, $\xi_3 = 2\sqrt{1 - \frac{1}{b}}$, for $b \geq 1$.

Consider a parameter $\mu \in D$ and $\bar{\mathbf{x}} = (\xi, f(\xi)) \in S_\mu$. The matrix associated with the linearized system around $\bar{\mathbf{x}}$ reads

$$M_\mu(\bar{\mathbf{x}}) = \begin{pmatrix} c(1 - \xi^2) & c \\ -\frac{1}{c} & -\frac{b}{c} \end{pmatrix}, \quad (7)$$

and the characteristic equation is

$$\lambda^2 - \lambda \left[c(1 - \xi^2) - \frac{b}{c} \right] - b(1 - \xi^2) + 1 = 0. \quad (8)$$

Denote by λ_1, λ_2 the corresponding eigenvalues.

The topological type of the hyperbolic equilibria is given by the signature of the determinant $\det M_\mu(\bar{\mathbf{x}}) = 1 - b(1 - \xi^2)$ and of the trace $\text{tr} M_\mu(\bar{\mathbf{x}}) = c(1 - \xi^2) - b/c$. Consider the sets Γ_1 and Γ_2 from the (b, a, c, x_1) space, defined by

$$\Gamma_1 = \{(\mu, x_1), |x_1| = \Gamma_1(b, c), a \geq 0, c > 0, b \leq c^2\}, \quad (9)$$

and

$$\Gamma_2 = \{(\mu, x_1), |x_1| = \Gamma_2(b), a \geq 0, c > 0, b \in (-\infty, 0) \cup [1, \infty)\}, \quad (10)$$

where $\Gamma_1(b, c) = \sqrt{1 - b/c^2}$ and $\Gamma_2(b) = \sqrt{1 - 1/b}$. With these notations we have $\det M_\mu(\bar{\mathbf{x}}) = 0$ if and only if $(\mu, \xi) \in \Gamma_2$ and $\text{tr} M_\mu(\bar{\mathbf{x}}) = 0$ if and only if $(\mu, \xi) \in \Gamma_1$.

In the non-hyperbolic case, if $\text{Re}(\lambda_1)$ and/or $\text{Re}(\lambda_2)$ vanish, nonlinear terms are to be considered in order to determine the topological type of $\bar{\mathbf{x}}$.

Sections in the sets Γ_1 and Γ_2 with a plane $c = \text{const.}$, $a = \text{const.}$ are given in Figures 2, 3, 4, for a $c < 1$, $c = 1$, and $c > 1$, respectively. The distinction was made by the number of intersections $\Gamma_1 \cap \Gamma_2$ in such a plane. Namely, for $c < 1$, $\Gamma_1 \cap \Gamma_2$ consists of two points (denoted by Q_1, Q_4), for $c = 1$, of three points (denoted by Q_1, Q_4, Q) and for $c > 1$ of four points ($Q_{1,2,3,4}$). In Figures 2, 3, 4, we have $\text{tr} M_\mu(\bar{\mathbf{x}}) > 0$ and $\det M_\mu(\bar{\mathbf{x}}) > 0$ in region I, $\det M_\mu(\bar{\mathbf{x}}) < 0$ in region II, and $\text{tr} M_\mu(\bar{\mathbf{x}}) < 0$ and $\det M_\mu(\bar{\mathbf{x}}) > 0$ in region III. Consequently, this corresponds to the regions where the sign of $\text{Re } \lambda_{1,2}$ is preserved.

Remark 3. In Fig. 2, equilibria $\bar{\mathbf{x}} = (\xi, f(\xi)) \in S_\mu$ with $(\mu, \xi) \in \Gamma_1$ situated between Q_1 and Q_4 have $\text{Re } \lambda_1 = \text{Re } \lambda_2 = 0$, $\lambda_{1,2} \in \mathbf{C} - \mathbf{R}$, so these equilibria are candidates for Hopf bifurcation. The same conclusion is valid for points situated between Q_1 and Q or Q_4 and Q on Γ_1 in Fig. 3, and for points between Q_1 and Q_2 or Q_3 and Q_4 on Γ_1 in Fig. 4.

Remark 4. In Figures 2, 3, 4, for equilibria $\bar{\mathbf{x}} = (\xi, f(\xi)) \in S_\mu$ with $(\mu, \xi) \in \Gamma_2$ we have $\det M_\mu(\bar{\mathbf{x}}) = 0$, hence at least one of the eigenvalues $\lambda_{1,2}$ is zero. Such equilibria are candidates for saddle-node bifurcation.

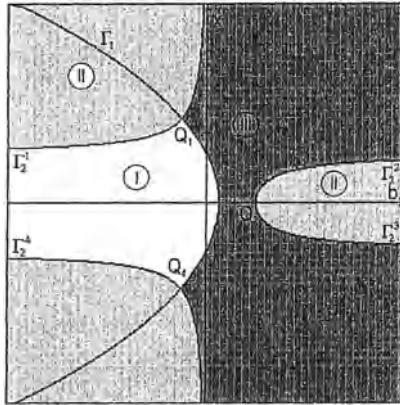


Figure 2. The regions in the (b, x_1) plane where the sign of $\text{Re } \lambda_{1,2}$ is preserved, if $c \in (0, 1)$: (I) $\text{Re } \lambda_1 > 0, \text{Re } \lambda_2 > 0$; (II) $\text{Re } \lambda_1 < 0, \text{Re } \lambda_2 > 0$, (III) $\text{Re } \lambda_1 < 0, \text{Re } \lambda_2 < 0$.

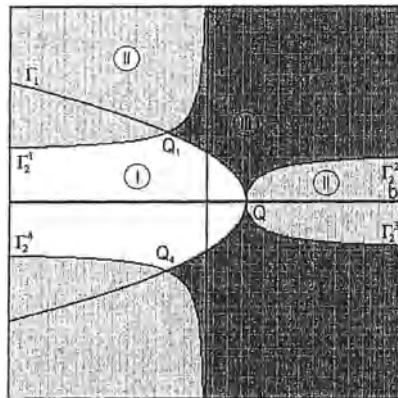


Figure 3. The regions in the (b, x_1) plane where the sign of $\text{Re } \lambda_{1,2}$ is preserved, if $c = 1$: (I) $\text{Re } \lambda_1 > 0, \text{Re } \lambda_2 > 0$; (II) $\text{Re } \lambda_1 < 0, \text{Re } \lambda_2 > 0$, (III) $\text{Re } \lambda_1 < 0, \text{Re } \lambda_2 < 0$.

Remark 5. Equilibria $\bar{x} = (\xi, f(\xi)) \in S_\mu$ with $(\mu, \xi) \in Q_i$, $i = 1, 4$, have $\lambda_1 = \lambda_2 = 0$, so they are candidates for a double zero bifurcation.

The projection of the set $\Sigma \cap \Gamma_1$ on the (b, a, c) space is the set $H_1 \cup H_2$, where

$$H_{1,2} : a = \pm \frac{1}{3c^3} (3c^2 - b^2 - 2bc^2) \sqrt{c^2 - b}. \quad (11)$$

Only the branches of Γ_1 with $b \in (-c, c)$ for $c > 1$ or $b \in (-c, c^2]$ for $c \leq 1$ separates regions with different signs of $\text{Re } \lambda_{1,2}$, so only these branches will be considered in the following.

The projection of $\Sigma \cap \Gamma_2$ on the (b, a, c) space is the set $S_1 \cup S_2$, given by (7).

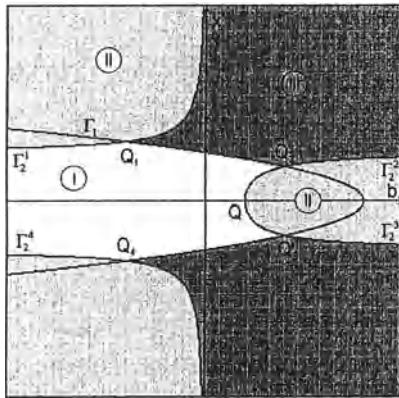


Figure 4. The regions in the (b, x_1) plane where the sign of $\text{Re } \lambda_{1,2}$ is preserved, if $c > 1$: (I) $\text{Re } \lambda_1 > 0, \text{Re } \lambda_2 > 0$; (II) $\text{Re } \lambda_1 < 0, \text{Re } \lambda_2 > 0$, (III) $\text{Re } \lambda_1 < 0, \text{Re } \lambda_2 < 0$.

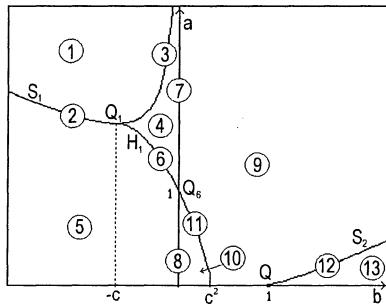


Figure 5. Section in the surfaces $H_{1,2} \cup S_{1,2}$ with a $c = \text{const.}$ plane, $c \in (0, 1)$.

The intersection of $H_1 \cup H_2$ with a $c = \text{const.}$ plane depends significantly on c . We found three typical cases, corresponding to $c < 1$, $c = 1$, $c > 1$, respectively. They are qualitatively illustrated in Figures 5, 6, 7. The curves of intersection of the surfaces H_1 , H_2 , S_1 , S_2 with a $c = \text{const.}$ plane where denoted in the same way as the corresponding surfaces. In every plane $c = \text{const.}$, at the intersection points Q_i , $i = 1, 4$, the curves $H_1 \cup H_2$ and $S_1 \cup S_2$ cuts tangentially. Consider also the points $Q_6 = H_1 \cap S_0$ and, for $c > 1$, $Q_0 = H_1 \cap H_2$, $Q_5 = H_1 \cap S_2$.

Taking into account the sign of $\text{Re } \lambda_{1,2}$ established in Figures 2, 3, 4, in Figures 5, 6, 7 the type of hyperbolic equilibria is emphasized. Thus, we have:

- for (b, a) situated in region 1: a saddle \times ;
- for (b, a) situated in region 4: two saddles and an attractor \times, \bullet, \times ;
- for (b, a) situated in region 5: two saddles and a repulsor \times, \circ, \times ;
- for (b, a) situated in region 7, 9: an attractor \bullet ;

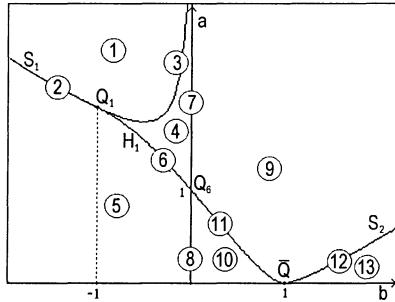


Figure 6. Section in the surfaces $H_{1,2} \cup S_{1,2}$ with a $c = \text{const.}$ plane, $c = 1$.

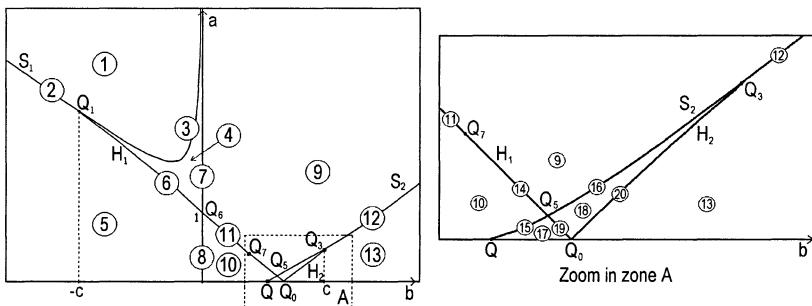


Figure 7. Section in the surfaces $H_{1,2} \cup S_{1,2}$ with a $c = \text{const.}$ plane, $c > 1$.

- for (b, a) situated in region 8, 10: a repulsor \circ ;
- for (b, a) situated in region 13: two attractors and a saddle \bullet, \times, \bullet ;
- for (b, a) situated in region 17: two repulsors and a saddle \circ, \times, \circ ;
- for (b, a) situated in region 18: a repulsor, a saddle and an attractor \circ, \times, \bullet .

In these regions the system possesses only hyperbolic equilibria. We also found the topological type (hyperbolic or not) for the other regions. Namely, we have:

- for (b, a) situated in the regions 2, 3 or at Q_1 : a saddle and a non-hyperbolic equilibrium \times, \square ;
- for (b, a) situated at Q_5 : two non-hyperbolic equilibria \square, \square ;
- for (b, a) situated in region 6: two saddles and a non-hyperbolic equilibrium \times, \square, \times ;
- for (b, a) situated in regions 11, 14, at Q_6 or at Q : a non-hyperbolic equilibrium \square ;
- for (b, a) situated in region 15: a non-hyperbolic equilibrium and a repulsor \square, \circ ;

- for (b, a) situated in regions 12, 16 or at Q_3 : a non-hyperbolic equilibrium and an attractor \square, \bullet ;
- for (b, a) situated at Q_0 : two non-hyperbolic equilibria and a saddle \square, \times, \square ;
- for (b, a) situated in region 19: a repulsor, a saddle and a non-hyperbolic equilibrium \circ, \times, \square ;
- for (b, a) situated in region 20: a non-hyperbolic equilibrium, a saddle and an attractor \square, \times, \bullet .

The topological type of the non-hyperbolic equilibria is established in Section 2.

2. Local bifurcation

For each non-hyperbolic equilibria we found the topological normal form. This allowed us to establish the topological type of non-hyperbolic equilibria and the local bifurcation generated by their presence.

2.1. Saddle-node bifurcation

Consider $\mu \in S_1 \cup S_2 - \{(b, a, c), b = \pm c\}$ and $\bar{\mathbf{x}} = (\xi, f(\xi)) \in S_\mu$ with $\xi^2 = 1 - \frac{1}{b}$. Thus, for the equilibrium $\bar{\mathbf{x}}$ we have $\lambda_1 = \frac{c^2 - b^2}{bc} \neq 0$ and $\lambda_2 = 0$

First, assume $b \neq 1$. Hence $\xi \neq 0$. Using certain changes of coordinates [7], [11], system (3) is topologically equivalent with the following system

$$\begin{cases} \dot{y}_1 = \lambda_1 y_1 + \frac{2c^2 \xi}{\lambda_1} y_1 y_2 + O(|\mathbf{y}|^3), \\ \dot{y}_2 = -\frac{\xi b^2}{\lambda_1} y_2^2 + O(|\mathbf{y}|^3), \end{cases} \quad (12)$$

which is the normal form for a non-degenerated saddle-node equilibrium. The bifurcation corresponding to such parameters is a codimension-one saddle-node bifurcation. The local dynamics around such an equilibrium is deduced using the center manifold theorem. As the center manifold for system (12) is given by $y_1 = 0$, up to third order terms, the flow on the center manifold is determined by

$$\dot{y}_2 = -\frac{\xi b^2}{\lambda_1} y_2^2 + O(|y_2|^3).$$

Thus, for $b \in (-\infty, -c) \cup (\min\{1, c\}, c)$ the equilibrium $\bar{\mathbf{x}}$ possesses three repulsive directions and an attractive one, while for $b \in (-c, 0) \cup (\max\{1, c\}, \infty)$ the equilibrium $\bar{\mathbf{x}}$ possesses three attractive directions and a repulsive one.

If $b = 1$ then $a = 0$ and $\xi = 0$ is the abscissa of the triple equilibrium. In this case, if $c \neq 1$ we have $\lambda_1 \neq 0$. Using appropriate transformations

[7], [12], system (3) is written as

$$\begin{cases} \dot{y}_1 = \lambda_1 y_1 - \frac{c^2}{\lambda_1} y_1 y_2^2 + O(|\mathbf{y}|^4), \\ \dot{y}_2 = \frac{1}{3\lambda_1} y_2^3 + O(|\mathbf{y}|^4), \end{cases} \quad (13)$$

which represents the topological normal form for a simple degenerated saddle-node equilibrium. The bifurcation corresponding to these parameters is a cusp bifurcation.

The flow on the center manifold $y_1 = 0$ is given by [10]

$$\dot{y}_2 = \frac{1}{3\lambda_1} y_2^3 + O(|\mathbf{y}|^4).$$

Therefore, the equilibrium $\bar{\mathbf{x}}$ is a weakly attractive degenerated saddle-node if $c < 1$ and a weakly repulsive degenerated saddle-node if $c > 1$.

2.2. Hopf bifurcation

Consider $\mu \in H_1 \cup H_2$ such that $b \in (-c, c^2]$ as $c < 1$ and $b \in (-c, c)$ as $c \geq 1$. In these cases the F-N system possesses an equilibrium $\bar{\mathbf{x}} = (\xi, f(\xi)) \in S_\mu$ with $\xi^2 = 1 - \frac{b}{c^2}$, for which $\lambda_{1,2} \in \mathbf{C} - \mathbf{R}$ and $\operatorname{Re}\lambda_1 = \operatorname{Re}\lambda_2 = 0$.

Using a sequence of transformations and making the computations in the complex field [7], system (3) is reduced, around a point \mathbf{x} near the equilibrium $\bar{\mathbf{x}}$, to its normal form:

$$\frac{dw}{d\theta} = (\gamma(\alpha) + i)w + l_1(\alpha)w^2\bar{w} + l_2(\alpha)w^3\bar{w}^2 + O(|w|^6), \quad (14)$$

where $l_1(\alpha)$ and $l_2(\alpha)$ are the first and the second Liapunov coefficients, respectively, and were computed in [7], [9]. For $\alpha = \mu$ and $\mathbf{x} = \bar{\mathbf{x}}$ we have $l_1 = \frac{-b^2+2bc^2-c^2}{2c\sqrt{1-\frac{b^2}{c^2}}^3}$ and $l_2 = \frac{10b^2(b-c^2)}{144c\sqrt{1-\frac{b^2}{c^2}}^3}$ if $l_1 = 0$ [11].

Note that $l_1 = 0$ iff $b = b^* = c^2 - c\sqrt{c^2 - 1}$. Consequently, (i) if $c \leq 1$, or $c > 1$ and $b \neq b^*$, then $l_1 \neq 0$, hence $\bar{\mathbf{x}}$ is a non-degenerated Hopf equilibrium and a codimension-one Hopf bifurcation takes place; (ii) if $c > 1$ and $b = b^*$ then $l_1 = 0$ and $l_2 \neq 0$, hence $\bar{\mathbf{x}}$ is a degenerated Hopf equilibrium of order two.

Denote by $Q_7(b^*, a^*, c)$ the corresponding point of H_1 . In the last case, a Bautin bifurcation takes place and from the point Q_7 emerge the curves of non-hyperbolic limit cycle bifurcation values, which is approximated asymptotically, near Q_7 , by the curve Ba defined by the equation

$$l_2 l_1^2 + 4 \frac{\operatorname{Re}\lambda}{\operatorname{Im}\lambda} = 0$$

deduced in [7], [9].

2.3. Bogdanov-Takens bifurcation

Consider $\mu \in (H_1 \cup H_2) \cap (S_1 \cap S_2)$, with $|b| = c$, and $\bar{x} = (\xi, f(\xi)) \in S_\mu$ with $\xi^2 = 1 - \frac{1}{b}$. The corresponding points in Figures 4, 5, 6 are denoted by Q_{1-4} . In this case $\lambda_1 = \lambda_2 = 0$.

Using appropriate transformations, the normal form of (3) at the point μ reads

$$\begin{cases} \dot{y}_1 = y_2, \\ \dot{y}_2 = -b\xi y_1^2 - 2c\xi y_1 y_2 - \frac{b}{3}y_1^3 - cy_1^2 y_2. \end{cases} \quad (15)$$

If $b \neq 1$, then $\xi \neq 0$, hence \bar{x} is a non-degenerated Bogdanov-Takens equilibrium. Around such an equilibrium, system (3) is topologically equivalent with the following

$$\begin{cases} \dot{y}_1 = y_2, \\ \dot{y}_2 = \beta_1 + s\beta_2 y_2 + y_1^2 - y_1 y_2, \end{cases} \quad (16)$$

where the coefficients $\beta_1 = -\frac{16c^4\bar{\xi}}{b^3}[a + (b-1)\bar{\xi} - b\bar{\xi}^3/3]$, $\beta_2 = 2(c^2/b^2 - 1)$ are given in [11] and $s = -1$ at Q_1, Q_4 , and $s = 1$ at Q_2, Q_3 , while $\bar{\xi} = s\sqrt{(1-1/b)}$. Consequently, the F-N system (3) exhibits at Q_{1-4} codimension-two Bogdanov-Takens bifurcations. In addition, in every plane $c = \text{const.}$, at the points Q_i emerge the curves HL_i corresponding to homoclinic bifurcation values [8], [11]. The equations for $HL_{1,3}$ are approximated by

$$HL_\pm : a = \frac{1.47(1 - b^2/c^2)^2 - 2(b-1)^2}{3b\bar{\xi}}.$$

If $b = 1$, then $\xi = 0$ and \bar{x} is a degenerated Bogdanov-Takens equilibrium of order two. Around this equilibrium, system (3) is topologically equivalent with the following

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = \varepsilon_1 + \varepsilon_2 x_1 + \varepsilon_3 x_2 - x_1^3 - x_1^2 x_2, \end{cases} \quad (17)$$

where

$$\varepsilon_1 = \frac{9\sqrt{3}c^3a}{b^2\sqrt{b}}, \varepsilon_2 = \frac{9c^2(b-1)}{b^2}, \varepsilon_3 = \frac{3(c^2-b)}{b}. \quad (18)$$

Consequently, a codimension three bifurcation takes place around this point.

2.4. Double-zero with symmetry of order two bifurcation

In the following we restrict to the case $a = 0$. Consider $\mu = (1, 0, 1)$. Then the F-N system (3) possesses a triple equilibrium $\bar{x} = (0, 0)$.

Around (μ, \bar{x}) , with μ situated in the $a = 0$ plane, system (3) is topologically equivalent with the following [13]

$$\begin{cases} \dot{y}_1 = y_2, \\ \dot{y}_2 = \varepsilon_2 y_1 + \varepsilon_3 y_2 - y_1^3 - y_1^2 y_2, \end{cases} \quad (19)$$

which is a normal form for a double zero with symmetry of order two bifurcation. Denote by $\bar{Q}(1, 1)$ the point corresponding to μ in the $a = 0$ plane. Consequently, from \bar{Q} the following bifurcation curves emerge [11]:

- $HL = \{(b, c), b > 1, \frac{3}{5b^2} [7c^2b - 12c^2 + 5b^2] + O(\varepsilon_2^{3/2}) = 0\}$, corresponding to double homoclinic bifurcation values,
- $B = \{(b, c), b > 1, \frac{3}{b^2} [(c_0 - 1)c^2b - c_0c^2 + b^2] + O(\varepsilon_2^{3/2}) = 0\}$, where $c_0 \approx 0.752$, corresponding to non-hyperbolic limit cycle bifurcation values,
- $H = \{(b, c), b > 1, 2c^2b - 3c^2 + b^2 = 0\}$, corresponding to two simultaneously Hopf bifurcation values,
- $R^+ = \{(b, c), b = 1, c^2 - b > 0\}$, $R^- = \{(b, c), b = 1, c^2 - b < 0\}$, corresponding to cusp bifurcation values,
- $H' = \{(b, c), b < 1, c^2 - b = 0\}$, corresponding to Hopf bifurcation values.

3. Concluding remarks

The analysis in Section 2 allowed us to completely determine the topological type of non-hyperbolic equilibria emphasized in Section 1.

Thus, the non-hyperbolic equilibria corresponding to parameter values situated on $S_1 \cup S_2$, in regions 2, 15, 16 in Figures 5, 6, 7, are partially repulsive saddle-nodes, while for parameter values situated in regions 3, 12, they are partially attractive saddle-nodes.

The non-hyperbolic equilibria corresponding to parameter values situated on $H_1 \cup H_2$, in regions 6, 11 or at Q_6 in Figures 5, 6, 7, are slowly attractive Hopf equilibria, while for parameter values situated in regions 14, 18, 20, they are slowly repulsive Hopf equilibria.

The non-hyperbolic equilibria corresponding to parameter values situated at Q_1 or Q_3 are non-degenerated Bogdanov-Takens equilibria, while for parameter values situated at Q_7 correspond degenerated Hopf (Bautin) equilibria.

In addition, non-generic codimension-two bifurcations take place at Q_5 and Q_0 due to the coexistence of two different non-hyperbolic equilibria, namely a partially repulsive saddle-node and a slowly repulsive Hopf equilibrium at Q_5 , and two slowly repulsive Hopf equilibria at Q_0 .

Finally, the non-hyperbolic equilibria corresponding to parameter values situated at Q are degenerated saddle-node (cusp) equilibria, which are slowly attractive for $c < 1$ and slowly repulsive for $c > 1$. For $c = 1$, the point $Q = \bar{Q}$ corresponds to degenerated Bogdanov-Takens of order two equilibrium, consequently to a codimension-three bifurcation. This is the unique generic codimension-three local bifurcation exhibited by the FitzHugh-Nagumo model (3).

Acknowledgment

The first author was partially supported by the grant MEC 260 44/2001.

References

- [1] Barnes B., Grimshaw R. *Analytical and numerical studies on the Bonhoeffer van der Pol system*, J. Aust. Math. Soc. Series B, **38**, (1995), 427-453.
- [2] Carr J. *Application of Center Manifold Theory*, Springer, New York, (1981).
- [3] Chow, S.N., Li, C., Wang, D. *Normal forms and bifurcation of Planar Vector Fields*, Cambridge University Press, 1994.
- [4] FitzHugh, R. *Impulses and physiological states in theoretical models of nerve membrane*, Biophysical J., **1**, 1961, pp. 445-466.
- [5] Hodgkin A.L., Huxley A.F. *A quantitative description of membrane current and its application to conduction and excitation in nerve*, J. Physiol., **117**, (1952), 500-544.
- [6] Kuznetsov, Yu. *Elements of applied bifurcation theory*, Springer, New York, 1995.
- [7] Rocsoreanu, C., Georgescu, A., Giurgiteanu, N., *The FitzHugh-Nagumo Model. Bifurcation and Dynamics*, Kluwer Academic Publishers, Dordrecht, 2000.
- [8] Rocsoreanu C., Giurgiteanu, Georgescu A., *Degenerated Hopf bifurcation in the FitzHugh-Nagumo system. 1. Bogdanov-Takens bifurcation*, Ann. Univ. Timisoara, **XXXV**, (2), 1997, pp.285-298.
- [9] Rocsoreanu C., Giurgiteanu, Georgescu A., *Degenerated Hopf bifurcation in the FitzHugh-Nagumo system. 1. Bautin bifurcation*, Revue D'analyse numérique et de théorie de l'approximation, **29**, 1, 2000, pp.97-109.
- [10] Sterpu, M., Georgescu A., *Codimension three bifurcations for the FitzHugh-Nagumo system*, Mathematical Reports, **3** (53),3, 2001 (in print).
- [11] Sterpu, M., *Dynamics and bifurcation for two generalized Van der Pol models*, Univ. of Pitești Publisher, 2001 (romanian).
- [12] Sterpu, M, *Saddle-node bifurcation in the FitzHugh-Nagumo Model*, Mathematical Reports, **4** (54), 3, 2002 (in print).
- [13] Sterpu, M, *A double-zero with symmetry of order two bifurcation in the FitzHugh-Nagumo model* Ann. of Univ. of Craiova, **XXIX**, 2002 (in print).

OPTIMIZATION OF STEADY-STATE FLOW OF INCOMPRESSIBLE FLUIDS *

Tomáš Roubíček

*Mathematical Institute, Charles University, Sokolovská 83, CZ-186 75 Praha 8, and
Institute of Information Theory and Automation, Academy of Sciences, Pod vodárenskou
věží 4, CZ-182 08 Praha 8, Czech Republic.*

tomas.roubicék@mff.cuni.cz

Abstract An optimal boundary-control problem for the steady-state buoyancy-driven flow of an incompressible, heat conductive fluid is investigated as far as optimality conditions and an increment formula concerns. The controlled system covers, in particular, the Oberbeck-Boussinesq model. Regularity of the Navier-Stokes system coupled with heat equation and its adjoint system is substantially used.

Keywords: Buoyancy-driven flow, optimality conditions, increment formula.

1. Introduction

Optimal control problem of steady-state incompressible flow coupled with heat equation was already studied in [2,4,8,9,19,20]. The isothermal case (i.e. without heat equation) was studied, e.g., in [1,3,5,10,15,16,17,27] and [22; Section III.11]. Due to generality of the controlled system (1) below, this contribution covers a lot of these results as a special case.

Besides, optimal control of evolutionary Navier-Stokes system was treated in a lot of other works, see e.g. monographs [11,12,22] and references therein.

2. The controlled system

Our a bit academical problem can be interpreted, after necessary simplifications, as a flow of, say, water heated up from outside to prepare a

*This research was partly covered by the grants 201/00/0768 (GA ČR), A 107 5005 (GA AV ČR), and MSM 11320007 (MŠMT ČR).

morning tee in a completely closed container Ω . Anyhow, one can think, e.g., about transport in a high-pressure vapor reactor [19,20] or flow a melted steel in a mold during a casting process, too.

We consider Ω a bounded smooth domain in \mathbb{R}^d , $d \leq 3$. To cover various possibilities, we consider the following fairly general system:

$$(u \cdot \nabla)u - \nu\Delta u + \nabla p = g(1 - \alpha_0\theta), \quad (1a)$$

$$\operatorname{div} u = 0 \quad (1b)$$

$$u \cdot \nabla\theta - \kappa\Delta\theta = \alpha_1|\nabla u|^2 + \alpha_2\theta g \cdot u, \quad (1c)$$

where u is the *velocity* vector field, p the *pressure*, θ *temperature*, and g an external (e.g. gravity) force. The material parameters are: ν the *viscosity*, κ heat *conductivity*, α_0 linearized relative mass density variation with respect to temperature, α_1 reflects *dissipation* effects, while α_2 expresses *adiabatic heat* effects. For a rigorous derivation of a system like (1) we refer to Kagei, Růžička and Thäter [21; System (16)] or Rajagopal, Růžička and Srinivasa [25]: it is shown that the coefficient α_1 depends on Ostrach's dissipation number, while the coefficient α_2 depends also on the Reynolds and the Prandtl numbers. The conventional *Oberbeck-Boussinesq model* uses $\alpha_1 = \alpha_2 = 0$. For derivation of the model with $\alpha_1 > 0$ but $\alpha_2 = 0$ see [24; Chap.I].

The system should be completed by boundary conditions. For simplicity, we will consider no-slip boundary condition for velocity and Newton's boundary condition for temperature, i.e.

$$u = 0, \quad \kappa \frac{\partial\theta}{\partial n} + b\theta = h \quad \text{on } \Gamma, \quad (2)$$

with n being the unit outward normal to the boundary $\Gamma := \partial\Omega$ of Ω .

We call the pair $(u, \theta) \in W_{0,\operatorname{DIV}}^{1,2}(\Omega; \mathbb{R}^d) \times W^{1,2}(\Omega)$ a weak solution to the boundary-value problem (1)–(2) if

$$\begin{aligned} & ((u \cdot \nabla)u, v) + \nu(\nabla u, \nabla v) + \kappa(\nabla\theta, \nabla\tilde{v}) + (u \cdot \nabla\theta, \tilde{v}) \\ & + (b\theta, \tilde{v})_\Gamma - ((1 - \alpha_0\theta)g, v) - (\alpha_1|\nabla u|^2 + \alpha_2\theta g \cdot u, \tilde{v}) \\ & - (h, \tilde{v})_\Gamma = 0 \quad \forall v \in W_{0,\operatorname{DIV}}^{1,2}(\Omega; \mathbb{R}^d), \quad \forall \tilde{v} \in W^{1,2}(\Omega), \end{aligned} \quad (3)$$

where $W_{0,\operatorname{DIV}}^{1,2}(\Omega; \mathbb{R}^d) := \{v \in W_0^{1,2}(\Omega; \mathbb{R}^d); \operatorname{div} v = 0\}$, (\cdot, \cdot) denotes the scalar product in $L^2(\Omega)$ or $L^2(\Omega; \mathbb{R}^d)$ or $L^2(\Omega; \mathbb{R}^{d \times d})$ and $(\cdot, \cdot)_\Gamma$ is the scalar product in $L^2(\Gamma)$. Moreover, in what follows we will denote by $\|\cdot\|_p$ the norm of $L^p(\Omega)$ or $L^p(\Omega; \mathbb{R}^d)$ or $L^p(\Omega; \mathbb{R}^{d \times d})$ and by $\|\cdot\|_{p,\Gamma}$ the norm of $L^p(\Gamma)$. Besides, N_Γ will be the norm of the trace operator $W^{1,2}(\Omega) \rightarrow L^2(\Gamma)$, and N_p and $N_{p,q}$ will denote the norm of the embedding $W_0^{1,2}(\Omega) \subset L^p(\Omega)$ and $L^p(\Omega) \subset L^q(\Omega)$, respectively.

The existence of a weak solution to the whole problem (1)-(2) is not automatic unless $\alpha_1 = \alpha_2 = 0$. Its uniqueness is even more delicate because, due to the quadratic terms on the right-hand side of (1c), we lack any global a-priori estimate if $\alpha_1 \neq 0 \neq \alpha_2$ and therefore we can get the uniqueness only of those solutions whose energy does not exceed certain limits.

For some $h_{\min} < h_{\max}$, we denote the set of admissible h 's in (2) by

$$\mathfrak{H}_{\text{ad}} = \{h \in L^\infty(\Gamma); h_{\min} \leq h(x) \leq h_{\max} \text{ for a.a. } x \in \Gamma\}. \quad (4)$$

Proposition 1 *Let $\varrho > N_2 N_\Gamma c_P^{-1} \max(|h_{\min}|, |h_{\max}|) \sqrt{\text{meas}_{d-1}(\Gamma)}$ and $\alpha_0, h_{\min}, h_{\max} \in \mathbb{R}$ be arbitrary, where $c_P = c_P(\kappa, b_{\min}, \Omega)$ is the constant from the Poincaré inequality $c_P \|\theta\|_{W^{1,2}(\Omega)}^2 \leq \kappa \|\nabla \theta\|_2^2 + b_{\min} \|\theta\|_{2,\Gamma}^2$. Let the assumptions*

$$\alpha_0 N_2 N_4^2 \|g\|_\infty \varrho < \nu, \quad (5a)$$

$$g \in L^\infty(\Omega; \mathbb{R}^d) \text{ has a potential, i.e. } g = \nabla \varphi, \quad (5b)$$

$$\alpha_1 \geq 0, \alpha_2 \geq 0 \text{ sufficiently small,} \quad (5c)$$

$$b_{\max} \geq b(x) \geq b_{\min} > 0 \text{ for a.a. } x \in \Gamma \quad (5d)$$

be satisfied. Then, for each $h \in \mathfrak{H}_{\text{ad}}$, there exists a weak solution (u, θ) to (1)-(2) satisfying $\|\theta\|_2 \leq \varrho$. Moreover, if

$$\nu, \kappa, b_{\min} \text{ are sufficiently large,} \quad (6)$$

this solution is determined uniquely and the mapping $h \mapsto (u, \theta) : L^2(\Omega) \rightarrow W^{1,2}(\Omega; \mathbb{R}^{d+1})$, restricted on \mathfrak{H}_{ad} , is Lipschitz continuous and (weak, norm)-continuous.

Proof. For the existence, we use a Schauder fixed-point argument similarly (but not entirely the same) as in [26]: we take $\vartheta \in L^2(\Omega)$, $\|\vartheta\|_2 \leq \varrho$, find a unique $u \in W_{0,\text{DIV}}^{1,2}(\Omega; \mathbb{R}^d)$ solving (in the weak sense) the Navier-Stokes system

$$(u \cdot \nabla) u - \nu \Delta u + \nabla p = g(1 - \alpha_0 \vartheta), \quad \text{div } u = 0, \quad u|_\Gamma = 0, \quad (7)$$

which is possible thanks to the assumption (5a). Testing (3) by $(v, \tilde{v}) := (u, 0)$ (note that the term $(g, v) = (\nabla \varphi, v) = -(\varphi, \text{div } v) = 0$ in (3) vanishes due to (5b)), the basic a-priori estimate

$$\|\nabla u\|_2 \leq \frac{N_2}{\nu} \|g\|_\infty \alpha_0 \varrho =: C_1 \quad (8)$$

is easily obtained. Then we take the unique $\theta \in W^{1,2}(\Omega)$ solving (in the weak sense) the heat equation

$$\begin{aligned} u \cdot \nabla \theta - \kappa \Delta \theta &= f, \quad \kappa \frac{\partial \theta}{\partial n} + b\theta|_{\Gamma} = h, \\ \text{with } f &\equiv f(u, \vartheta) := \alpha_1 |\nabla u|^2 + \alpha_2 \vartheta g \cdot u, \end{aligned} \tag{9}$$

The (nowadays standard) regularity result

$$\|u\|_{\infty} \leq c \|u\|_{W^{2,2}(\Omega; \mathbb{R}^d)} \leq C_2 \equiv C_2(\Omega, \varrho) \tag{10}$$

is known, see e.g. [7] or [14; Chap.VIII, Thm.5.2]. This shows, in particular, that $f \in L^2(\Omega)$ so that the unique weak solution to (9) certainly does exist. This regularity together with the compact embedding $W^{2,2}(\Omega; \mathbb{R}^d) \subset W^{1,6-\varepsilon}(\Omega; \mathbb{R}^d)$ shows that the mapping $\vartheta \mapsto f(u(\vartheta), \vartheta) : L^2(\Omega) \rightarrow L^2(\Omega)$ is (weak,weak)-continuous; the limit passage in the term $\alpha_2 \vartheta g \cdot u$ is easy because, due to (10) and compactness of the embedding $W^{2,2}(\Omega) \subset L^\infty(\Omega)$, the mapping $\vartheta \mapsto u : L^2(\Omega) \rightarrow L^\infty(\Omega; \mathbb{R}^d)$ is (weak,norm)-continuous. By standard arguments, $f \mapsto \theta : L^2(\Omega) \rightarrow W^{1,2}(\Omega)$ is (weak,norm)-continuous. Altogether, $\vartheta \mapsto \theta : L^2(\Omega) \rightarrow L^2(\Omega)$ is (weak,norm)-continuous. Note that, if $\alpha_1 = \alpha_2 = 0$, then

$$c_{\Gamma} \|\theta\|_2 \leq N_2 N_{\Gamma} \max(|h_{\min}|, |h_{\max}|) \sqrt{\text{meas}_{d-1}(\Gamma)}. \tag{11}$$

If ϱ is taken as assumed and (5c) holds, the mapping $\vartheta \mapsto \theta$ maps the ball $\{\vartheta \in L^2(\Omega); \|\vartheta\|_2 \leq \varrho\}$ into itself and, by Schauder's theorem, it has a fixed point θ . Then $\theta = \vartheta$ together with the corresponding $u = u(\vartheta)$ from (7) solves (1)-(2).

Now, we will show the claimed Lipschitz-continuous dependence on h of this fixed point and, as a by-product, we get its uniqueness as well as its (weak,norm)-continuity. Write the integral identity (3) for two right-hand sides, say h_1 and h_2 . Then (u_1, θ_1) and (u_2, θ_2) will denote (some of) the corresponding solutions. Now we subtract these identities (3) and test it by $v := u_1 - u_2$ and $\tilde{v} = \theta_1 - \theta_2$. Then, abbreviating shortly $u_{12} \equiv u_1 - u_2$, $\theta_{12} \equiv \theta_1 - \theta_2$, and $h_{12} \equiv h_1 - h_2$ and using the identities

$$\begin{aligned} ((u_1 \cdot \nabla) u_1 - (u_2 \cdot \nabla) u_2, u_{12}) &= ((u_1 \cdot \nabla) u_{12}, u_{12}) \\ + ((u_{12} \cdot \nabla) u_2, u_{12}) &= ((u_{12} \cdot \nabla) u_2, u_{12}) \end{aligned} \tag{12}$$

and

$$\begin{aligned} (u_1 \cdot \nabla \theta_1 - u_2 \cdot \nabla \theta_2, \theta_{12}) &= (u_1 \cdot \nabla \theta_{12}, \theta_{12}) \\ + (u_{12} \cdot \nabla \theta_2, \theta_{12}) &= (u_{12} \cdot \nabla \theta_2, \theta_{12}) \end{aligned} \tag{13}$$

(which hold thanks to $\operatorname{div} u_1 = 0$ and $u_1|_\Gamma = 0$) and also

$$(|\nabla u_1|^2 - |\nabla u_2|^2, \theta_{12}) = (\nabla(u_1 + u_2), (\nabla u_{12})\theta_{12}) \quad \text{and} \quad (14)$$

$$(\theta_{1g} \cdot u_1 - \theta_{2g} \cdot u_2, \theta_{12}) = (\theta_{1g} \cdot u_{12}, \theta_{12}) + (g \cdot u_2, \theta_{12}^2), \quad (15)$$

we get, by testing (3) by $(v, \tilde{v}) := (u_{12}, \theta_{12})$,

$$\begin{aligned} & \nu \|\nabla u_{12}\|_2^2 + \kappa \|\nabla \theta_{12}\|_2^2 + (b\theta_{12}, \theta_{12})_\Gamma \\ &= (h_{12}, \theta_{12})_\Gamma - ((u_{12} \cdot \nabla) u_2, u_{12}) - (u_{12} \cdot \nabla \theta_2, \theta_{12}) \\ & \quad - \alpha_0(\theta_{12}, g \cdot u_{12}) + \alpha_1(\nabla(u_1 + u_2), (\nabla u_{12})\theta_{12}) \\ & \quad + \alpha_2(\theta_{1g} \cdot u_{12}, \theta_{12}) + \alpha_2(g \cdot u_2, \theta_{12}^2) \\ &\leq N_\Gamma \|h_{12}\|_2 \|\theta_{12}\|_{W^{1,2}(\Omega)} + \|\nabla u_2\|_2 \|u_{12}\|_4^2 + \|u_{12}\|_4 \|\nabla \theta_2\|_2 \|\theta_{12}\|_4 \\ & \quad + \alpha_0 \|\theta_{12}\|_2 \|g\|_\infty \|u_{12}\|_2 + \alpha_1 N_{6,4}^2 \|\nabla(u_1 + u_2)\|_6 \|\nabla u_{12}\|_2 \|\theta_{12}\|_6 \\ & \quad + \alpha_2 \|\theta_1\|_6 \|g\|_\infty \|u_{12}\|_6 \|\theta_{12}\|_6 + \alpha_2 N_{6,3}^2 \|g\|_\infty \|u_2\|_\infty \|\theta_{12}\|_2^2. \end{aligned} \quad (16)$$

Using Young and Poincaré inequalities, all right-hand-side terms but $\|h_{12}\|_2^2$ can be absorbed in the left-hand side as (6) is assumed. ■

Remark 2 The above uniqueness proof suggests a modification of the existence proof by showing a Lipschitz continuity of the mapping $\vartheta \mapsto \theta$ and then to use Banach's fixed-point theorem instead of the Schauder one. The requirement of contractivity of $\vartheta \mapsto \theta$ would, however, require (6) quantitatively stronger.

Remark 3 Usually, we are given by ν , κ , and b_{\min} . Then the choice of a suitable ϱ needs further analysis and (6) needs, in particular, a very small Reynolds number, cf. [2] for quantitative estimates.

3. The optimal control problem

We will consider the velocity/temperature tracking problem. For the purpose of Proposition 10 below, we consider two different norms for velocity, distinguished by $\xi = 0, 1$, and then the optimal-control problem:

$$\left\{ \begin{array}{ll} \text{Minimize} & J(u, \theta, h) := \int_\Omega \frac{\xi}{2} |\nabla u - \nabla u_d|^2 + \frac{\xi-1}{2} |u - u_d|^2 \\ & + \frac{1}{2} |\theta - \theta_d|^2 dx + \int_\Gamma h^2 dS \quad (\text{cost functional}) \\ \text{subject to} & (u, \theta, h) \in W_{0,\text{DIV}}^{1,2}(\Omega; \mathbb{R}^d) \times W^{1,2}(\Omega) \times L^\infty(\Gamma) \\ & \text{solves (1)–(2) weakly,} \quad (\text{state system}) \\ & \|\theta\|_2 \leq \varrho, \quad (\text{"selection" criterion}) \\ & h_{\min} \leq h(x) \leq h_{\max} \text{ for } x \in \Gamma, \quad (\text{control constraints}) \end{array} \right.$$

where u_d and θ_d are the desired velocity and temperature profiles. Let us emphasize that (\mathfrak{P}) is not a linear/quadratic optimization problem because the controlled system (1), containing 4 bilinear/quadratic terms, is obviously nonlinear.

Remark 4 (*The selection criterion.*) The criterion $\|\theta\|_2 \leq \varrho$ uses ϱ from Proposition 1 and is not a state constraint in the usual sense because it does not constrain implicitly the controls but it only wants to avoid a possible nonuniqueness in the response. This selection criterion can be omitted for the Oberbeck-Boussinesq system (i.e. if $\alpha_1 = \alpha_2 = 0$).

Proposition 5 *Let u_d belong to $L^2(\Omega; \mathbb{R}^d)$ (if $\xi = 0$) or to $W^{1,2}(\Omega; \mathbb{R}^d)$ (if $\xi = 1$), $\theta_d \in L^2(\Omega)$, and the assumptions (5)–(6) be satisfied. Then (\mathfrak{P}) has a solution.*

Proof. The set \mathfrak{H}_{ad} of admissible controls is compact in weak L^2 -topology in which the state mapping $h \mapsto (u(h), \theta(h))$, $\|\theta\|_2 \leq \varrho$, is continuous and the cost functional J lower-semicontinuous. Hence the minimum of $\Phi(h) = J(u(h), \theta(h), h)$ does exist by Bolzano-Weierstraß' theorem. ■

4. Optimality conditions

Of course, since the controlled system is nonlinear, there may exists (beside the globally optimal control whose existence has been claimed in Proposition 5) also locally optimal controls. Let us consider a fixed locally optimal reference pair $(\bar{u}, \bar{\theta}, \bar{h})$. We begin with the first-order optimality conditions. Formally, they can be found by applying the well-known Lagrange principle, where the state-equations are eliminated by the *Lagrange function*

$$\begin{aligned} L(u, \theta, h, w, \vartheta) &= J(u, \theta, h) - ((u \cdot \nabla)u, w) - \nu(\nabla u, \nabla w) \\ &\quad - \kappa(\nabla \theta, \nabla \vartheta) - (u \cdot \nabla \theta, \vartheta) + (h - b\theta, \vartheta)_\Gamma \\ &\quad + ((1 - \alpha_0 \theta)g, w) + (\alpha_1 |\nabla u|^2 + \alpha_2 \theta g \cdot u, \vartheta), \end{aligned} \quad (17)$$

cf. (3). Obviously, for fixed multipliers $w \in W_{0,\text{DIV}}^{1,2}(\Omega; \mathbb{R}^d)$ and $\vartheta \in W^{1,2}(\Omega)$, the function $L(\cdot, \cdot, \cdot, w, \vartheta) : W_{0,\text{DIV}}^{1,2}(\Omega; \mathbb{R}^d) \times W^{1,2}(\Omega) \times L^2(\Gamma) \rightarrow \mathbb{R}$ is quadratic and continuous, hence it is a C^2 -function, too. According to the Lagrange principle, $(\bar{u}, \bar{\theta}, \bar{h})$ should satisfy the necessary optimality conditions for minimizers of L with respect to $h \in \mathfrak{H}_{ad}$, i.e.

$$[L'_u(\bar{u}, \bar{\theta}, \bar{h}, w, \vartheta)](u) = 0 \quad \forall u \in W_{0,\text{DIV}}^{1,2}(\Omega; \mathbb{R}^d), \quad (18a)$$

$$[L'_\theta(\bar{u}, \bar{\theta}, \bar{h}, w, \vartheta)](\theta) = 0 \quad \forall \theta \in W^{1,2}(\Omega), \quad (18b)$$

$$[L'_h(\bar{u}, \bar{\theta}, \bar{h}, w, \vartheta)](h - \bar{h}) \geq 0 \quad \forall h \in \mathfrak{H}_{ad}. \quad (18c)$$

The identities (18a,b) lead to the *adjoint system* which has (in the classical formulation) the form:

$$(\nabla \bar{u})^\top w - (\bar{u} \cdot \nabla)w - \nu \Delta w + \nabla \pi = 2\alpha_1 \operatorname{div}(\vartheta \nabla \bar{u}) + \alpha_2 \bar{\theta} g \vartheta \quad (19a)$$

$$+ \bar{\theta} \nabla \vartheta + \begin{cases} \bar{u} - u_d & \text{for } \xi = 0 \\ \Delta(u_d - \bar{u}) & \text{for } \xi = 1 \end{cases}$$

$$\operatorname{div} w = 0, \quad (19b)$$

$$\kappa \Delta \vartheta + \bar{u} \nabla \vartheta + \alpha_2 g \cdot \bar{u} \vartheta = \alpha_0 g \cdot w + \theta_d - \bar{\theta} \quad (19c)$$

with the boundary conditions

$$w = 0, \quad \kappa \frac{\partial \vartheta}{\partial n} + b \vartheta = 0 \quad \text{on } \Gamma, \quad (20)$$

for the so-called *adjoint velocity, pressure, and temperature* $(w, \pi, \vartheta) \in W_{0, \operatorname{DIV}}^{1,2}(\Omega; \mathbb{R}^d) \times L^2(\Omega) \times W^{1,2}(\Omega)$ associated with a given state $(\bar{u}, \bar{\theta})$. Note that $(\nabla u)^\top w - (u \cdot \nabla)w$ means $(\sum_{k=1}^n (\frac{\partial u_k}{\partial x_i} w_k - u_k \frac{\partial w_i}{\partial x_k}))_{i=1, \dots, n}$.

Proposition 6 (1st-order necessary conditions.) *Let (5)–(6) hold, and let \bar{h} be a locally optimal control for (\mathfrak{P}) with associated state $(\bar{u}, \bar{\theta})$. Then the variational inequality*

$$\bar{h}(x) = \begin{cases} h_{\max} & \text{if } -\vartheta(x) > 2h_{\max}, \\ -\vartheta(x)/2 & \text{if } -\vartheta(x) \in [2h_{\min}, 2h_{\max}], \\ h_{\min} & \text{if } -\vartheta(x) < 2h_{\min}, \end{cases} \quad (21)$$

is satisfied for $\vartheta = \vartheta(\bar{u}, \bar{\theta}) \in W^{1,2}(\Omega)$ being, together with $w = w(\bar{u}, \bar{\theta})$, the unique weak solution to the adjoint system (19)–(20).

Proof. (Sketched.) In view of the special form (4) of $\mathfrak{H}_{\operatorname{ad}}$, (18c) is equivalent to (21) and, as already told, (18a,b) is equivalent to the adjoint system (19)–(20) whose solution can be shown to exist under the assumption (5)–(6). ■

To perform 2nd-order analysis (as, in the context of fluid control used in [10, 18, 27]), we need 2nd-order derivative of the Lagrange function. In view of (17), the 2nd differential of $L(\cdot, \cdot, \cdot, w, \vartheta)$ at a point (u, θ, h) is

$$\begin{aligned} L''_{(u, \theta, h)}(u, \theta, h, w, \vartheta)[(u_1, \theta_1, h_1), (u_2, \theta_2, h_2)] &= -((u_1 \cdot \nabla)u_2, w) \quad (22) \\ &\quad - ((u_2 \cdot \nabla)u_1, w) - (u_1 \cdot \nabla \theta_2, \vartheta) - (u_2 \cdot \nabla \theta_1, \vartheta) \\ &\quad + 2\alpha_1(\vartheta \nabla u_1, \nabla u_2) + \dot{\alpha}_2(\theta_1 g \cdot u_2, \vartheta) + \alpha_2(\theta_2 g \cdot u_1, \vartheta) \\ &\quad + (\theta_1, \theta_2) + (h_1, h_2)_\Gamma + \begin{cases} (u_1, u_2) & \text{if } \xi = 0, \\ (\nabla u_1, \nabla u_2) & \text{if } \xi = 1. \end{cases} \end{aligned}$$

This quadratic form is therefore independent of (u, θ, f) and bounded provided the multipliers w and ϑ are bounded in L^∞ -norm, as indeed shown in the proof of Proposition 10 below under the assumption b smooth and, for $\gamma = 2$,

$$\theta_d \in L^\gamma(\Omega), \quad u_d \in \begin{cases} L^\gamma(\Omega; \mathbb{R}^d) & \text{if } \xi = 0, \\ W^{2,\gamma}(\Omega; \mathbb{R}^d) & \text{if } \xi = 1. \end{cases} \quad (23)$$

By using Green formula and $\operatorname{div} \tilde{u} = 0$ and $\tilde{u}|_\Gamma = 0$, (22) restricted on the diagonal $(u_1, \theta_1, h_1) = (u_2, \theta_2, h_2)$ becomes symmetric and takes the form

$$\begin{aligned} L''_{(u,\theta,h)}(u, \theta, h, w, \vartheta)(\tilde{u}, \tilde{\theta}, \tilde{h})^2 &= 2((\tilde{u} \cdot \nabla)w, \tilde{u}) + 2(\tilde{u} \cdot \nabla \vartheta, \tilde{\theta}) \\ &+ 2\alpha_1(\vartheta \nabla \tilde{u}, \nabla \tilde{u}) + 2\alpha_2(\tilde{\theta}g, \tilde{u}\vartheta) + \|\tilde{h}\|_{2,\Gamma}^2 + \|\tilde{\theta}\|_2^2 + \begin{cases} \|\tilde{u}\|_2^2 & \text{if } \xi = 0, \\ \|\nabla \tilde{u}\|_2^2 & \text{if } \xi = 1. \end{cases} \end{aligned} \quad (24)$$

Proposition 7 (2nd-order sufficient conditions.) *Let (5)–(6) be valid, b smooth, (23) hold with $\gamma = 2$, and let $(\bar{u}, \bar{\theta}, \bar{h}, w, \vartheta)$ satisfy (1)–(2), $\bar{h} \in \mathfrak{H}_{\text{ad}}$, the first-order necessary conditions (19)–(20), together with the second-order sufficient condition:*

$$L''_{(u,\theta,h)}(\bar{u}, \bar{\theta}, \bar{h}, w, \vartheta)(u, \theta, h)^2 \geq \delta \|h\|_{2,\Gamma}^2 \quad (25)$$

for $L''_{(u,\theta,h)}$ from (24), some $\delta > 0$, and for all (u, θ, h) solving the system (1)–(2) linearized at $(\bar{u}, \bar{\theta}, \bar{h})$, i.e. for (u, θ, h) satisfying (1b), (2), and

$$(u \cdot \nabla)\bar{u} + (\bar{u} \cdot \nabla)u - \nu \Delta u + \nabla p = -g\alpha_0\theta, \quad (26a)$$

$$u \cdot \nabla \bar{\theta} + \bar{u} \cdot \nabla \theta - \kappa \Delta \theta = -2\alpha_1(\nabla \bar{u}, \nabla u) + \alpha_2 g \cdot (\theta \bar{u} + \bar{\theta}u) \quad (26b)$$

in the weak sense. Then $(\bar{u}, \bar{\theta}, \bar{h})$ is locally optimal with respect to the topology of $W^{1,2}(\Omega; \mathbb{R}^{d+1}) \times L^2(\Gamma)$.

Proof. (Sketched.) By [6], (25)–(26) yields $\Phi''(\bar{h})(h, h) \geq \delta_1 \|h\|_{2,\Gamma}^2$ for some $\delta_1 > 0$ and for all $h \in \mathfrak{H}_{\text{ad}}$. Moreover, one can prove that $(u, \theta, h, w, \vartheta) \mapsto L''_{(u,\theta,h)}(u, \theta, h, w, \vartheta)$ is continuous, and this continuity is inherited also by $\Phi''(\cdot)(h, h)$, so that one can conclude that \bar{h} is locally optimal for Φ with respect to the norm of $L^2(\Gamma)$. By the continuity of the state mapping $h \mapsto (u, \theta) : L^2(\Gamma) \rightarrow W^{1,2}(\Omega; \mathbb{R}^{d+1})$, see Proposition 1, this gives the claimed local optimality of $(\bar{u}, \bar{\theta}, \bar{h})$. ■

5. The increment formula

For $h \in \mathfrak{H}_{\text{ad}}$, as in the proof of Proposition 5, we denote $\Phi(h) = J(u(h), \theta(h), h)$ with $\|\theta(h)\|_2 \leq \varrho$. Recall that such $\theta = \theta(h)$ and $u = u(h)$ are unique under the assumption (5)–(6).

For the following lemma, let us abbreviate $z := (u, \theta) \in Z := W_{0,\text{DIV}}^{1,2}(\Omega; \mathbb{R}^d) \times W^{1,2}(\Omega)$ and $H = L^2(\Gamma)$, and define $\Pi : Z \times H \rightarrow Z^*$ by $\langle \Pi(z, h), \tilde{z} \rangle$ = the left-hand side of (3); thus $\Pi(z, h) = 0$ is just (3).

Lemma 8 (Increment formula for bi-quadratic problems.) *Let $J : Z \times H \rightarrow \mathbb{R}$ and $\Pi : Z \times H \rightarrow Z^*$ be quadratic mappings and $\lambda \in Z^{**} \cong Z$ satisfy $L'_z(z, h, \lambda) = 0$ with $L(z, h, \lambda) = J(z, h) - \lambda \circ \Pi(z, h)$, and let $\Pi(z, h) = 0$ and $\Pi(\tilde{z}, \tilde{h}) = 0$. Then*

$$\Phi(\tilde{h}) - \Phi(h) - \Phi'(h)(\tilde{h} - h) = \frac{1}{2}L''_{(z,h)}(z, h, \lambda)(\tilde{z} - z, \tilde{h} - h)^2. \quad (27)$$

Proof. As $L(\cdot, \cdot, \lambda)$ is quadratic, the Taylor expansion up to 2nd-order term is exact, i.e.

$$\begin{aligned} L(\tilde{z}, \tilde{h}, \lambda) &= L(z, h, \lambda) + L'_{(z,h)}(z, h, \lambda)(\tilde{z} - z, \tilde{h} - h) \\ &\quad + \frac{1}{2}L''_{(z,h)}(z, h, \lambda)(\tilde{z} - z, \tilde{h} - h)^2. \end{aligned} \quad (28)$$

As $\Pi(z, h) = 0$, $L(z, h, \lambda) = J(z, h) - \lambda \circ \Pi(z, h) = \Phi(h)$. Similarly, $\Pi(\tilde{z}, \tilde{h}) = 0$ implies $L(\tilde{z}, \tilde{h}, \lambda) = \Phi(\tilde{h})$. Moreover, using the adjoint equation $L'_z(z, h, \lambda) = 0$, we get

$$\begin{aligned} L'_{(z,h)}(z, h, \lambda)(\tilde{z} - z, \tilde{h} - h) &= L'_z(z, h, \lambda)(\tilde{z} - z) + L'_h(z, h, \lambda)(\tilde{h} - h) \\ &= J'_h(z, h)(\tilde{h} - h) - [\Pi'_h(z, h)(\tilde{h} - h)]^* \lambda = \Phi'(h)(\tilde{h} - h). \end{aligned} \quad \blacksquare \quad (29)$$

As the controlled system (1) as well as the cost functional in (3) are quadratic, we can use Lemma 8 for our problem:

Corollary 9 *The following increment formula for (3) holds:*

$$\begin{aligned} J(\tilde{u}, \tilde{\theta}, \tilde{h}) - J(u, \theta, h) - (\vartheta + 2h, \tilde{h} - h)_\Gamma &= ((\tilde{u} - u) \cdot \nabla w, \tilde{u} - u) + ((\tilde{u} - u) \cdot \nabla \vartheta, \tilde{\theta} - \theta) \\ &\quad + \alpha_1(\vartheta(\nabla(\tilde{u} - u), \nabla(\tilde{u} - u)) + \alpha_2((\tilde{\theta} - \theta)g, (\tilde{u} - u)\vartheta) \\ &\quad + \frac{1}{2}\|\tilde{h} - h\|_{2,\Gamma}^2 + \frac{1}{2}\|\tilde{\theta} - \theta\|_2^2 + \begin{cases} \frac{1}{2}\|\tilde{u} - u\|_2^2 & \text{if } \xi = 0, \\ \frac{1}{2}\|\nabla(\tilde{u} - u)\|_2^2 & \text{if } \xi = 1. \end{cases} \end{aligned} \quad (30)$$

where $\tilde{u} = u(\tilde{f})$, $u = u(f)$, $w = w(u, \theta)$, and $\vartheta = \vartheta(u, \theta)$.

Proof. Use (27) with L'' from (24) and $\lambda = (w, \vartheta)$, and realize that $\Phi'(h) = L'_h(u, \theta, h, w, \vartheta) = \vartheta + 2h$. \blacksquare

The formula (30) can be used for global analysis of (3) provided one shows the multipliers w and ϑ sufficiently small in the $W^{1,\infty}$ -norm.

Proposition 10 Let (5)–(6) be valid, b smooth, (23) hold for $\gamma > d$, let $\xi = 1$ or $\alpha_1 = 0$, and, for any $h \in \mathfrak{H}_{\text{ad}}$, the corresponding adjoint variables (w, ϑ) satisfy

$$\frac{1}{2} \geq A_{w,\vartheta} := \begin{cases} \|\nabla w\|_\infty + \frac{1}{2}\|\nabla \vartheta\|_\infty + \frac{1}{2}\alpha_2\|g\|_\infty\|\vartheta\|_\infty & \text{if } \xi = 0, \\ N_2^2(\|\nabla w\|_\infty + \frac{1}{2}\|\nabla \vartheta\|_\infty + \frac{1}{2}\alpha_2\|g\|_\infty\|\vartheta\|_\infty) + \alpha_1\|\vartheta\|_\infty & \text{if } \xi = 1, \end{cases} \quad (31a)$$

$$\frac{1}{2} \geq A_\vartheta := \frac{1}{2}\|\nabla \vartheta\|_\infty + \frac{1}{2}\alpha_2\|g\|_\infty\|\vartheta\|_\infty. \quad (31b)$$

Then Φ is strictly convex on \mathfrak{H}_{ad} .

Proof. First, let us prove $W^{1,\infty}$ -estimates of the multipliers w and ϑ . The adjoint equation (19) can equally be viewed as the Stokes system and the Poisson equation with the right-hand sides

$$\begin{aligned} f_1 &= -(\bar{u} \cdot \nabla)\bar{w} + (\nabla \bar{u})^\top \bar{w} + 2\alpha_1 \nabla \vartheta \cdot \nabla \bar{u} + 2\alpha_1 \vartheta \Delta \bar{u} \\ &\quad + \alpha_2 \bar{\theta} g \cdot \vartheta - \bar{\theta} \nabla \vartheta + \begin{cases} \bar{u} - u_d & \text{for } \xi = 0 \\ \Delta(u_d - \bar{u}) & \text{for } \xi = 1 \end{cases} \end{aligned} \quad (32)$$

$$f_2 = \bar{u} \nabla \vartheta + \alpha_2 g \cdot \bar{u} \vartheta + \bar{\theta} - \theta_d. \quad (33)$$

By the proof of Proposition 1, we know that $\bar{u} \in W^{2,2}(\Omega; \mathbb{R}^d)$ and $\bar{\theta} \in W^{1,2}(\Omega)$, so that $f := (\bar{u} \cdot \nabla)\bar{u} + \alpha_0 g \theta \in L^6(\Omega; \mathbb{R}^d)$ and, by the $W^{2,p}$ -regularity [14] of Stokes' system with the right-hand side f , we get $\bar{u} \in W^{2,\gamma}(\Omega; \mathbb{R}^d)$; as $n \leq 3$ we may assume $\gamma \leq 6$. Besides, certainly $w \in W_{0,\text{DIV}}^{1,2}(\Omega; \mathbb{R}^d)$ and $\vartheta \in W^{1,2}(\Omega)$. As $W^{1,2}(\Omega) \subset L^6(\Omega)$ and $W^{2,\gamma}(\Omega) \subset W^{1,\infty}(\Omega)$, we can see that $f_2 \in L^2(\Omega)$ due to the estimate

$$\|f_2\|_2 \leq \|\bar{u}\|_\infty \|\nabla \vartheta\|_2 + \alpha_2 N_{6,2} \|g\|_\infty \|\bar{u}\|_\infty \|\vartheta\|_6 + \|\bar{\theta} - \theta_d\|_2. \quad (34)$$

By the $W^{2,2}$ -regularity of the equation $\kappa \Delta \vartheta + f_2 = 0$ with the boundary conditions $\kappa \partial \vartheta / \partial n + b \vartheta = 0$ (here we need b smooth), we can see that $\vartheta \in W^{2,2}(\Omega)$. Then we have also $f_1 \in L^2(\Omega; \mathbb{R}^d)$ due to the estimate

$$\begin{aligned} \|f_1\|_2 &\leq \|\bar{u}\|_\infty \|\nabla \bar{w}\|_2 + N_{6,4}^2 \|\nabla \bar{u}\|_6 \|\bar{w}\|_6 + 2\alpha_1 N_{6,4}^2 \|\nabla \vartheta\|_6 \|\nabla \bar{u}\|_6 \\ &\quad + 2\alpha_1 \|\vartheta\|_\infty \|\Delta \bar{u}\|_2 + \alpha_2 N_{\infty,2} \|\bar{\theta}\|_\infty \|g\|_\infty \|\vartheta\|_\infty \\ &\quad + N_{6,2} \|\bar{\theta}\|_\infty \|\nabla \vartheta\|_6 + \begin{cases} N_{\infty,2} \|\bar{u} - u_d\|_\infty & \text{for } \xi = 0 \\ \|\Delta(u_d - \bar{u})\|_2 & \text{for } \xi = 1 \end{cases} \end{aligned} \quad (35)$$

Again by regularity for Stokes' systems [14; Chap. IV, Thm.6.1], we have $w \in W^{2,2}(\Omega; \mathbb{R}^d)$. By usual bootstrap argument, we can see that f_2 then belongs even to $L^\gamma(\Omega)$ (recall that $\gamma \leq 6$) due to the estimate

$$\begin{aligned} \|f_2\|_\gamma &\leq N_{6,\gamma} \|\bar{u}\|_\infty \|\nabla \vartheta\|_6 \\ &\quad + \alpha_2 N_{\infty,\gamma} \|g\|_\infty \|\bar{u}\|_\infty \|\vartheta\|_\infty + \|\bar{\theta} - \theta_d\|_\gamma. \end{aligned} \quad (36)$$

Hence, we get one of the desired estimates: $\vartheta \in W^{2,\gamma}(\Omega) \subset W^{1,\infty}(\Omega)$. Then, by the obvious “bootstrap” modification of (35), we have also $f_1 \in L^\gamma(\Omega; \mathbb{R}^d)$. Again by $W^{2,\gamma}$ -regularity for Stokes’ system [14], we have the remaining desired estimate $w \in W^{2,\gamma}(\Omega; \mathbb{R}^d) \subset W^{1,\infty}(\Omega; \mathbb{R}^d)$.

Then, in the case $\xi = 1$, we can estimate the terms with indefinite signs in (30) by the Hölder and Young inequalities as follows:

$$\begin{aligned} & ((\tilde{u} - u) \cdot \nabla) w, \tilde{u} - u + ((\tilde{u} - u) \cdot \nabla \vartheta, \tilde{\theta} - \theta) \quad (37) \\ & + \alpha_1 (\vartheta \nabla (\tilde{u} - u), \nabla (\tilde{u} - u)) + \alpha_2 ((\tilde{\theta} - \theta) g, (\tilde{u} - u) \vartheta) \\ & \geq -\|\tilde{u} - u\|_2^2 \|\nabla w\|_\infty - \|\tilde{u} - u\|_2 \|\nabla \vartheta\|_\infty \|\tilde{\theta} - \theta\|_2 \\ & - \alpha_1 \|\vartheta\|_\infty \|\nabla (\tilde{u} - u)\|_2^2 - \alpha_2 \|\tilde{u} - u\|_2 \|g\|_\infty \|\vartheta\|_\infty \|\tilde{\theta} - \theta\|_2 \\ & \geq -A_{w,\vartheta} \|\nabla (\tilde{u} - u)\|_2^2 - A_\vartheta \|\tilde{\theta} - \theta\|_2^2. \end{aligned}$$

Then (31) ensures that these terms are dominated by last two terms in (30) so that $\Phi(\tilde{h}) - \Phi(h) - \Phi'(h)(\tilde{h} - h) \geq 0$ by (27). Replacing the role of h and \tilde{h} , we get $\Phi(h) - \Phi(\tilde{h}) - \Phi'(\tilde{h})(h - \tilde{h}) \geq 0$, and by adding these inequalities we get monotonicity of Φ' , hence convexity of Φ . The term $\frac{1}{2} \|\tilde{h} - h\|_{2,\Gamma}^2$ in (30) then ensures strict convexity of Φ .

For $\alpha_1 = 0$, we can handle the case $\xi = 0$ by simplified estimation as suggested in (31a). ■

Corollary 11 (1st-order sufficient condition.) *Under the conditions in Proposition 10, (21) ensures that \bar{h} is the unique optimal control to (\mathfrak{P}) .*

Remark 12 The increment-formula technique has been developed in a general context in [13]. Increment formula (30) has been derived in [23] for mere Navier-Stokes system (in particular, the $W^{1,\infty}$ -regularity of w has been proved by Málek [23] for this case) and in [2] for Oberbeck-Boussinesq system (i.e. $\alpha_1 = \alpha_2 = 0$). Severe restrictions of conditions like (31) have been analyzed by Bubák [2].

References

- [1] N. Bilić: Approximation of optimal distributed control problem for Navier-Stokes equations. In: *Numerical Methods and Approx. Th.*, Univ. Novi Sad, Novi Sad, 1985, 177-185.
- [2] P. Bubák: Optimal control of flow driven by the thermal field. MS-diploma thesis, Math.-Phys. Faculty, Charles University, Prague, 2002.
- [3] J. Burkard, J. Peterson: Control of steady incompressible 2D channel flow. In: *Flow Control* (M.D. Gunzburger, ed.), IMA Vol. Math. Appl. **68**, Springer, New York, 1995, 111–126.
- [4] A. Capatina, R. Stavre: Optimal control of a non isothermal Navier-Stokes flow. *Int. J. Eng. Sci.* **34** (1996), 59-66.
- [5] E. Casas: Optimality conditions for some control problems of turbulent flow. In: *Flow Control* (M.D. Gunzburger, ed.), IMA Vol. Math. Appl. **68**, Springer, New York, 1995, 127-147.

- [6] E. Casas, F. Tröltzsch: Second order necessary and sufficient optimality conditions for optimization problems and applications to control theory, to appear in *SIAM J. Optimization*.
- [7] P. Constantin, P. Foias: *Navier-Stokes equations*. The University of Chicago Press, 1989.
- [8] C. Cuvelier: Optimal control of a system governed by the Navier-Stokes equations coupled with the heat equation. In: *New Dev. Differ. Equat.* (W.Eckhaus, ed.) North-Holland, 1976, 81-98.
- [9] C. Cuvelier: Resolution numerique d'un probleme de controle optimal d'un couplage des equations de Navier-Stokes et celle de la chaleur. *Calcolo* **15** (1980), 345-379.
- [10] M.C. Desai, K. Ito: Optimal control of Navier-Stokes equations. *SIAM J. Control Optim.* **32** (1994), 1428-1446.
- [11] H.O. Fattorini: *Infinite Dimensional Optimization and Control Theory*. Cambridge Univ. Press, Cambridge, 1999.
- [12] A.V. Fursikov: *Optimal Control of Distributed Systems. Theory and Applications*. AMS, Providence, 2000.
- [13] R. Gabasov, F. Kirillova: *The Qualitative Theory of Optimal Processes*. Marcel Dekker, New York, 1976 (Russian orig.: Nauka, Moscow, 1971).
- [14] P.G. Galdi: *An introduction to the Navier-Stokes Equations*. Springer-Verlag, 1994.
- [15] M.D. Gunzburger: A prehistory of flow control and optimization. In: *Flow Control* (M.D.Gunzburger, ed.), IMA Vol. Math. Appl. **68**, Springer, New York, 1995, 185-195.
- [16] M.D. Gunzburger, L. Hou, T.P. Svobodny: Analysis and finite element approximation of optimal control problems for stationary Navier-Stokes equations with distributed and Neumann controls. *Math. Comp.* **57** (1991), 123-151.
- [17] M.D. Gunzburger, L. Hou, T.P. Svobodny: Boundary velocity control of incompressible flow with an application to viscous drag reduction. *SIAM J. Control Optim.* **30** (1992), 167-181.
- [18] M. Hinze: A remark on second order methods in control of fluid flow. *Zeitschrift Angew. Math. Mech.* **81**, Suppl. 3, (2001).
- [19] K. Ito, J.S. Scroggs, H.T. Tran: Optimal control of thermally coupled Navier-Stokes equations. In: *Optimal design and control*. (J.Borggaard et al., eds.) Birkhäuser, Boston, 1995, 199-214.
- [20] K. Ito, H.T. Tran, J.S. Scroggs: Mathematical issues in optimal design of a vapor transport reactor. In: *Flow Control* (M.D.Gunzburger, ed.), IMA Vol. Math. Appl. **68**, Springer, New York, 1995, 197-218.
- [21] Y. Kagei, M. Růžička, G. Thäter: Natural Convection with Dissipative Heating. *Comm. Math. Physics*, **214** (2000), 287-313.
- [22] J.L. Lions: *Contrôle des systèmes distribués singuliers*. Bordas, Paris, 1983. Engl. transl.: *Control of Distributed Singular Systems*. Gauthier-Villars, 1985.
- [23] J. Málek, T. Roubíček: Optimization of steady flows for incompressible viscous fluids. In: *Nonlinear Applied Analysis*. (Eds. A.Sequiera, H. Beirão da Vega, J.H. Videman) Plenum Press, New York, 1999, 355-372.
- [24] B. Mohammadi, O. Pironneau: *Analysis of the K-Epsilon turbulence model*. J.Wiley, Chichester, 1994.
- [25] K.R. Rajagopal, M. Růžička, A.R. Srinivasa: On the Oberbeck-Boussinesq Approximation. *Math. Models Methods Appl. Sci.* **6** (1996), 1157-1167.
- [26] T. Roubíček: Steady-state buoyancy-driven viscous flow with measure data. *Mathematica Bohemica*, **126**, (2001), 493-504.
- [27] T. Roubíček, F. Tröltzsch: Lipschitz stability of optimal controls for the steady-state Navier-Stokes equations. *Control & Cybernetics*, submitted.

TOPOLOGICAL DEGREE APPROACH TO STEADY STATE FLOW

Cristina Sburlan

*Department of Mathematics, "Ovidius" University,
Bd. Mamaia 124, 8700-Constantza, Romania
c_sburlan@univ-ovidius.ro*

Silviu Sburlan

*Department of Mathematics, "Ovidius" University,
Bd. Mamaia 124, 8700-Constantza, Romania
ssburlan@univ-ovidius.ro*

Abstract In this paper we study the steady state flow equation, which can be treated as an eigenvalue problem, and we shall apply the results from the topological degree theory and from the bifurcation theory.

Keywords: Steady state flow, coincidence degree, bifurcation points

Let $\Omega \subseteq R^N$ ($\Omega = R^N - \mathcal{P}$), $2 \leq N \leq 3$, be a domain with enough smooth boundary, $Q = \Omega \times (0, +\infty)$, $\Sigma = \partial\Omega \times (0, +\infty)$, with \mathcal{P} a bounded region in R^N .

We consider the Navier-Stokes system, for the flow of an incompressible fluid:

$$(\nabla \cdot u)(x, t) = 0 \quad (\text{the incompressibility condition}) \quad (1)$$

$$u_t(x, t) + (u \cdot \nabla)(x, t) - \nu \Delta u(x, t) = \nabla p(x, t) + f(x, t), \quad (x, t) \in Q \quad (2)$$

(the Navier – Stokes flow equation)

$$u = 0 \text{ on } \Sigma \text{ and } u \rightarrow (1, 0, 0) \text{ for } |x| \rightarrow +\infty. \quad (3)$$

We suppose that the body forces are of potential type, *i.e.*

$$f(x, t) = \nabla_x V(x, t) \quad (4)$$

and we note $q := p + V$, where p is the (unknown) pressure in fluid. Then we can write (2) under the form:

$$u_t(x, t) + (u \cdot \nabla) u(x, t) - \nu \Delta u(x, t) = \nabla q(x, t), \quad (5)$$

where ν is the dynamical viscosity ($\nu = \frac{1}{Re}$ is the inverse of the Reynolds number). Here, the velocity $u = (u_1, \dots, u_N)$ and the "pressure" q are not known, and they must be determined from the system (1) – (3).

We shall study the case of the steady state flow: $u_t = 0 \Leftrightarrow u = \text{const.}$ The equation (5) becomes:

$$(u \cdot \nabla) u(x, t) - \nu \Delta u(x, t) = \nabla q(x, t) \quad (6)$$

Let $X := \{y \in (L^2(\Omega))^N; \nabla \cdot y = 0, y \cdot n = 0 \text{ on } \partial\Omega\}$ be the Hilbert space of "incompressible fluids" and $E := \{y \in (H_0^1(\Omega))^N; \nabla \cdot y = 0\}$ be a subspace of X , and $P : (L^2(\Omega))^N \rightarrow X$ is Leray projector.

Denote by $A \in L(E, E)$ the Stokes operator:

$$(Ay, w) = \sum_{i=0}^n \int_{\Omega} \nabla y_i \cdot \nabla w_i dx, \forall y, w \in E$$

and define the nonlinear form:

$$b(y, z, w) := \sum_{i,j=1}^N \int_{\Omega} y_i D_i z_j w_j dx$$

which determines the nonlinear operator $C : E \rightarrow E$:

$$C(y, w) := b(y, y, w), \forall y, w \in E.$$

Then we can reformulate the problem (6) as the problem

$$\nu Ay + C(y) = P(q) \quad (7)$$

where A is simmetric, i.e. $(Ay, w) = (y, Aw)$, and strongly monotone, because $(Ay, y) \geq \|y\|^2$. This is an eigenvalue problem, having ν as eigenvalue parameter.

Because the embedding $E \hookrightarrow X$ is compact ($2 \leq N \leq 3$) (from extension of Sobolev theorem for unbounded domains – see [2]), we have that the operator $\widehat{C} : E \subset X \rightarrow X$, $\widehat{C} = I \circ C$, is compact (the composition of a compact operator with a continuous one). Similary, the operator $\widetilde{A} = I \circ A$ is compact, too.

Using the fact that $\widehat{C}(y, w)$ is linear in w , we can prove that there exists $c > 0$ such that:

$$\|\widehat{C}(y)\| \leq c \cdot \|y\|, \forall y \in E, \quad (8)$$

where c is the supremum of the "pressure" of fluid.

Denote by $T : X \rightarrow X$, $T := P(q)$ and $\tilde{C} := \hat{C} - T$, and by (7), there exists $k > 0$ such that:

$$\|\tilde{C}(y)\| \leq k \cdot \|y\|, \forall y \in E, \quad (9)$$

The equation (7) can be written:

$$\nu \tilde{A}y + \tilde{C}(y) = 0 \quad (10)$$

Now, for two operators $\mathcal{L} : D(\mathcal{L}) \subseteq X \rightarrow X$ and $\mathcal{N} : D(\mathcal{N}) \subseteq X \rightarrow X$, where \mathcal{L} is linear and maximal monotone, and \mathcal{N} is compact, we have that: $\nu\mathcal{L} + \mathcal{N} = I + \nu\mathcal{L} - I + \mathcal{N} = (I + \nu\mathcal{L})(I - (I + \nu\mathcal{L})^{-1}(I - \mathcal{N}))$.

Because $\mathcal{L} : X \rightarrow X$ is maximal monotone, we have that $I + \nu\mathcal{L}$ is invertible, and the equation $\nu\mathcal{L}y + \mathcal{N}(y) = 0$ becomes:

$$(I - (I + \nu\mathcal{L})^{-1}(I - \mathcal{N}))(y) = 0 \quad (11)$$

Denote by $M(\nu) := (I + \nu\mathcal{L})^{-1}(I - \mathcal{N})$, and remark that this operator is a compact one.

If $D \subset X$ is an open bounded set such that: $\nu\mathcal{L}y + \mathcal{N}(y) \neq 0, \forall y \in \partial D$, then we define the coincidence degree of the pair $(\mathcal{L}, \mathcal{N})$, relatively to D by $d_\nu((\mathcal{L}, \mathcal{N}), D) := d_{LS}(I - M(\nu), D, 0)$, where d_{LS} denotes the Leray-Schauder degree. So, this coincidence degree has the properties of the Leray-Schauder degree.

Theorem 1 *If $d_\nu((\mathcal{L}, \mathcal{N}), D) \neq 0$ then the equation $\nu\mathcal{L}y + \mathcal{N}(y) = 0$ has at least one solution in D .*

Proof. We have that $d_{LS}(I - M(\nu), D, 0) \neq 0$, and from the solution property for the Leray-Schauder degree (see [1]) there exists $y \in D$ such that $(I - M(\nu))y = 0$, and so it results that $\nu\mathcal{L}y + \mathcal{N}(y) = 0$.

Theorem 2 *Let $(\nu_t)_{t \in [0,1]} \subset (0, \infty)$ be a continuous deformation such that the equation $\nu_t\mathcal{L}y + \mathcal{N}(y) = 0$ has no solutions $y \in \partial D$, for all $t \in [0, 1]$. Then $d_{\nu_t}((\mathcal{L}, \mathcal{N}), D)$ is independent of $t \in [0, 1]$.*

Proof. The conclusion results from the invariance of Leray-Schauder degree to the homotopy (see [1]).

Now, we want to find the bifurcation points of equation (10), i.e. the points $(\nu, 0)$ accumulating nontrivial solutions of (10).

Denote by $\mathcal{C}(\tilde{A})$ the set of all characteristic values of \tilde{A} . Because \tilde{A} is linear, the algebraic multiplicity of this characteristic values is 1.

We can state the following result:

Theorem 3 If $\nu_0 \in \mathcal{C}(\tilde{A})$ is such that $\text{dist}(\nu_0, \mathcal{C}(\tilde{A}) \setminus \{\nu_0\}) > 2$, then the equation (10) has at least one bifurcation point in $(\nu_0 - 1, \nu_0 + 1)$.

Proof. We have that the algebraic multiplicity of ν_0 , $m(\nu_0) = 1$.

$$\text{Let } \varepsilon = \text{dist}(\nu_0, \mathcal{C}(\tilde{A}) \setminus \{\nu_0\}) - 2 > 0, \nu_1 = \nu_0 - 1 - \varepsilon, \nu_2 = \nu_0 + 1 + \varepsilon.$$

$$\text{We have that: } \text{dist}(\nu_1, \mathcal{C}(\tilde{A})) = \text{dist}(\nu_2, \mathcal{C}(\tilde{A})) = 1 + \varepsilon.$$

We shall prove that for any $r > 0$, the equation (10) has a solution (ν_r, y_r) such that: $\nu_r \in (\nu_0 - 1, \nu_0 + 1)$ and $\|y_r\| = r$.

Let $r > 0$ be arbitrary fixed. Assume by contradiction that

$$\nu \tilde{A}y + \tilde{C}(y) \neq 0, \forall \|y\| = r \text{ and } \nu \in (\nu_1, \nu_2).$$

We have that

$$\begin{aligned} d_{\nu_1}((0, \tilde{C}), B(0, r)) &= d_{LS}(I - (I - \tilde{C}), B(0, r), 0) = \\ &= d_{\nu_2}((0, \tilde{C}), B(0, r)) \end{aligned} \quad (12)$$

Consider the homotopies $H_t^i = \nu_i t \left(\tilde{A} - \frac{1}{\nu_i} I \right) + (1-t) \tilde{C}$, $t \in [0, 1]$, $i = 1, 2$. Because $\tilde{A} : D(\tilde{A}) \subseteq X \rightarrow X$ is maximal monotone, the operator $\tilde{A} - \frac{1}{\nu_i} I$ is also maximal monotone. We have

$$\begin{aligned} \|\nu_i \tilde{A}y\| &\leq \|\nu_i \tilde{A}y - y + y\| \leq \|\nu_i \tilde{A}y - y\| + \|y\| \Rightarrow \\ \|\nu_i \tilde{A}y - y\| &\geq \|\nu_i \tilde{A}y\| - \|y\| \geq \nu_i \|y\| - \|y\| \geq \varepsilon \cdot \|y\| > 0, \forall y \in B(0, r), i=1, 2. \end{aligned}$$

By the invariance of the coincidence degree to homotopy we have:

$$d_{\nu_i} \left(\left(\tilde{A} - \frac{1}{\nu_i} I, 0 \right), B(0, r) \right) = d_{\nu_i} \left((0, \tilde{C}), B(0, r) \right), i = 1, 2. \quad (13)$$

We know that $\tilde{A}^{-1} : X \rightarrow X$ is linear, continuous and compact.

From (12) – (13) we obtain

$$d_{\nu_1} \left(\left(\tilde{A} - \frac{1}{\nu_1} I, 0 \right), B(0, r) \right) = d_{\nu_2} \left(\left(\tilde{A} - \frac{1}{\nu_2} I, 0 \right), B(0, r) \right),$$

i.e.

$$d_{LS} \left(I - \frac{1}{\nu_1} \tilde{A}^{-1}, B(0, r), 0 \right) = d_{LS} \left(I - \frac{1}{\nu_2} \tilde{A}^{-1}, B(0, r), 0 \right) \quad (14)$$

But

$$d_{LS} \left(I - \frac{1}{\nu_1} \tilde{A}^{-1}, B(0, r), 0 \right) = (-1)^{m(\nu_0)} d_{LS} \left(I - \frac{1}{\nu_2} \tilde{A}^{-1}, B(0, r), 0 \right)$$

and $m(\nu_0) = 1$, which contradicts (14).

Now, we have that

$$\nu\tilde{A} + \tilde{C} = I + \nu\tilde{A} - I + \tilde{C} = (I + \nu\tilde{A}) \left(I - (I + \nu\tilde{A})^{-1} (I - \tilde{C}) \right),$$

so, the equation (10) is equivalent to

$$\begin{aligned} & \left(I - (I + \nu\tilde{A})^{-1} (I - \tilde{C}) \right) (y) = 0 \Leftrightarrow \\ & y = (I + \nu\tilde{A})^{-1} (I - \tilde{C}) (y) \Leftrightarrow \\ & y = (I + \nu\tilde{A})^{-1} y + (I + \nu\tilde{A})^{-1} \tilde{C} (y) \end{aligned} \quad (15)$$

where the operator $(I + \nu\tilde{A})^{-1} \tilde{C}$ is compact, as the composition of a compact operator with a continuous one.

Denoting by $L := (I + \nu\tilde{A})^{-1}$ and by $N := (I + \nu\tilde{A})^{-1} \tilde{C}$, the equation (15) becomes

$$y = Ly + N(y), \quad (16)$$

where $\|N(y)\| \leq c \cdot \|y\|$, with $c > 0$.

It is easy to see that $\mu = 1$ is a characteristic value for \tilde{L} and its multiplicity is 1.

Consider now the eigenvalue problem:

$$y = \mu\tilde{L}y + N(y) \quad (17)$$

Denote by $\mathcal{C}_0 = \{[\mu, 0] ; \mu \in R\}$, \mathcal{S}_0 the set of nontrivial solutions, and $\mathcal{S} := \overline{\mathcal{S}_0}$.

For $\mu = 1$, the equation (17) is in fact the equation (16).

Based on the well-known results from the bifurcation theory (see [1]), namely Krasnoselskii and Rabinowitz theorems, we can state the following

Theorem 4 *The point $[1, 0] \in \mathcal{C}_0$ is a bifurcation point for equation (17).*

Proof. Because $\mu = 1$ is a characteristic value of \tilde{L} with odd multiplicity, we can apply Krasnoselskii theorem, and thus $[1, 0]$ is a bifurcation point for equation (17).

Theorem 5 *Under the above conditions, \mathcal{S} contains a connected component \mathcal{E} , passing by $[1, 0] \in \mathcal{C}_0$, with one of the properties:*

1. \mathcal{E} is unbounded in $R \times X$;
2. \mathcal{E} contains a finite number of points $[\mu_j, 0]$ with μ_j characteristic values of \tilde{L} . Moreover, the number of points with odd multiplicity – including $[1, 0]$ – is even.

Proof. Indeed, for $\mu = 1$, which has odd multiplicity, the result follows from Rabinowitz theorem.

Comments. Writing Navier-Stokes system as an eigenvalue problem in the case of steady state flow we can deduce not only the existence of the solutions, but also the branch structure of solutions and the bifurcation points. This is perfectly concordant with the physical meaning of steady state flow. Moreover, we see the power of topological methods for solving such problems.

References

- [1] S. Sburlan, *Gradul topologic. Lecții asupra ecuațiilor nelineare*, Ed. Academiei Române, București, 1983.
- [2] D. Pascali, S. Sburlan, *Nonlinear Mappings of Monotone Type*, Sijhoff & Noordhoff Int. Publ., 1978.
- [3] C. Sburlan, S. Sburlan, Abstract Fourier Method for Stokes Equation, Bull. PAMM-2039, 2002 (to appear).
- [4] S. Sburlan, C. Mortici, *A Coincidence Degree for Bifurcation Problems*, BAM-1784/2001 XCIV, 2000.
- [5] S. Sburlan, C. Mortici, *Topological Methods for Semilinear Problems with Maximal Monotone Nonlinearity*, Conference at 26th Congress of American Romanian Academy of Arts and Sciences, July 25-29, Montreal, 2001.
- [6] S. Sburlan, L. Barbu and C. Mortici, *Ecuatii Diferentiale, Integrale și Sisteme Dinamice*, Ed. Ex Ponto, Constanța, 1999.
- [7] S. Sburlan, G. Morosanu, *Monotonicity Methods for Partial Differential Equations*, MB-11/PAMM, Budapest, 1999.
- [8] R. Temam, *Navier-Stokes Equations*, North-Holland Publishing Company, 1977.

FAST NUMERICAL ALGORITHMS FOR WIENER SYSTEMS IDENTIFICATION *

Vasile Sima

*National Institute for Research & Development in Informatics,
Bd. Maresal Averescu 8-10, 71316 Bucharest 1
Romania
vsima@iciadmin.ici.ro*

Abstract A Wiener system consists of a linear dynamic block followed by a static nonlinearity. The identification of a Wiener system means finding a mathematical model using the input and output data. The approach chosen for identification uses a state space representation for the linear part and a single layer neural network to model the static nonlinearity. Fast subspace identification algorithms are used for estimating the linear part, based on the available input-output data. Using the resulted state-space model, an approximate model of the nonlinear part is found by an improved Levenberg-Marquardt (LM) algorithm. Finally, the whole model is refined using a specialized, MINPACK-like, but structure-exploiting LAPACK-based LM algorithm. The output normal form is used to parameterize the linear part. With a suitable ordering of the variables, the Jacobian matrices have a block diagonal form, with an additional block column at the right. This structure is preserved in a QR factorization with column pivoting restricted to each block column. The implementation is memory conserving and about one order of magnitude faster than standard LM algorithms or specialized LM calculations based on conjugate gradients for solving linear systems.

Keywords: Conjugate gradients, least-squares approximation, Levenberg-Marquardt algorithm, optimization, system identification.

1. Introduction

A discrete-time Wiener system has a state space representation

$$x(k+1) = Ax(k) + Bu(k),$$

*Work partially supported by the European Community BRITE-EURAM III *Thematic Networks Programme NICONET* (project BRRT-CT97-5040).

$$\begin{aligned} z(k) &= Cx(k) + Du(k), \\ y(k) &= f(z(k)) + v(k), \end{aligned} \quad (1)$$

where $x(k) \in \mathbb{R}^n$, $u(k) \in \mathbb{R}^m$, and $y(k) \in \mathbb{R}^\ell$ are the state, input, and output vectors at time k , respectively, $v(k)$ is a zero-mean stochastic disturbance term, A , B , C , and D are real matrices of appropriate dimensions, and $f(\cdot)$ is a nonlinear vector function from \mathbb{R}^ℓ to \mathbb{R}^ℓ . Briefly speaking, a Wiener system consists of a linear dynamic block followed by a static nonlinearity.

The *Wiener system identification problem* can be stated as follows: Given the input and output data sequences of the system (1), $\{u(i)\}$, and $\{y(i)\}$, $i = 1, \dots, N$, where the input sequence $\{u(k)\}$ is assumed sufficiently persistently exciting, and statistically independent from the perturbation $\{v(k)\}$, find: (a) the order n ; (b) an estimate of the quadruple (A, B, C, D) of the Wiener state space model and the initial conditions (up to a similarity transformation); (c) an approximation of $f(\cdot)$.

The approach chosen to solve the above identification problem uses a state space representation for the linear part and a single layer neural network to model the static nonlinearity. Fast subspace identification algorithms are used for estimating the linear part, based on the available input-output data. The resulted state-space model is used for finding an approximate model of the nonlinear part by a Levenberg-Marquardt (LM) algorithm. Finally, the whole model is refined using a specialized, MINPACK-like [6], [7], but structure-exploiting LAPACK-based [1] scaling-invariant LM algorithm. The output normal form is used to parameterize the linear part. The parameters corresponding to the nonlinear part come first in the global parameter vector. Using this ordering, the Jacobian matrices in the multi-output case ($\ell > 1$) have a block diagonal form, with an additional block column at the right. This structure is preserved in a QR factorization with column pivoting restricted to each block column, which makes sense in the identification context. The rank deficient case is also covered. Incremental condition estimation is optionally used for finding the ranks of matrices. The approach is implemented in a specialized toolbox of the freely available Subroutine Library In COntrol Theory (SLICOT). The Jacobian is computed analytically, for the nonlinear part, and numerically, for the linear part. The implementation is memory conserving and significantly faster than standard LM algorithms or specialized LM calculations based on conjugate gradients (without preconditioning) for solving linear systems.

A systematic approach for solving the Wiener system identification problem is given in [14], and further developed in Section 2. Easy-to-

use software components based on this approach are briefly described in Section 3. Finally, Section 4 presents part of the numerical results obtained using this software.

2. Algorithmic Outline

The Wiener identification problem can be solved in a systematic manner [14], [10]. Algorithmic improvements are summarized below.

Step 1: Estimating the linear part

Given the sequences $\{u(k), y(k)\}$, and assuming that the function $f(\cdot)$ contains odd terms when evaluating its Taylor series expansion, the linear time-invariant (LTI) part of the Wiener system is identified, using one of the subspace identification techniques [13]. If the static nonlinearity $f(\cdot)$ is even, a subspace solution has been provided in a similar way [15]. Therefore, it is possible to identify the linear dynamics as in case the nonlinearities were absent and extract information on the structure of the linear part.

The most time consuming calculation for this step is devoted for finding an upper triangular factor of a QR factorization for a large matrix built from two block-Hankel matrices (with input and output data). Several algorithmic options are available for performing these computations:

- Structure-exploiting correlation calculations and Cholesky factorization [11];
- Fast QR factorization [5];
- Standard QR factorization.

The first two approaches could be one-two orders of magnitude faster than the third approach.

Step 2: Estimating the parameters of $f(\cdot)$

The nonlinear part is modelled as a set of single layer neural networks,

$$f_s(z(k)) = \hat{f}_s(z(k)) + \epsilon_s(k), \quad s = 1, \dots, \ell, \quad (2)$$

$$\hat{f}_s(z(k)) := \sum_{i=1}^{\nu} \left(\alpha(s, i) \phi \left(\sum_{j=1}^{\ell} \beta(s, i, j) z_j(k) + b(s, i) \right) \right) + b(s, \nu + 1).$$

The vector $\epsilon(k)$ is the approximation error. The coefficients $\alpha(s, i)$, $\beta(s, i, j)$, $b(s, i)$ and $b(s, \nu + 1)$ are real numbers to be estimated, and

the integer ν represents the number of neurons. These constants are stacked in the parameter vector $\theta \in \mathbb{R}^{\ell((\ell+2)\nu+1)}$,

$$\theta = \left(\theta_1^T | \theta_2^T | \dots | \theta_\ell^T \right)^T := \\ (\beta(1, 1, 1), \dots, \beta(1, \nu, \ell), \alpha(1, 1), \dots, \alpha(1, \nu), b(1, 1), \dots, b(1, \nu + 1) \\ | \beta(2, 1, 1), \dots | \dots | \beta(\ell, 1, 1), \dots)^T,$$

and are estimated by solving the nonlinear least squares (NLS) problem

$$\min_{\theta} \sum_{k=1}^N \left\| \begin{bmatrix} y_1(k) - \hat{f}_1(\hat{z}(k)) \\ \vdots \\ y_\ell(k) - \hat{f}_\ell(\hat{z}(k)) \end{bmatrix} \right\|^2, \quad (3)$$

where $\|\cdot\|$ denotes the Euclidean norm, and the sequence $\{\hat{z}(k)\}_{k=1}^N$ is an estimated output sequence of the linear part, computed using estimates of the quadruple (A, B, C, D) determined in Step 1. Since the parameters appearing in a row of the vector in the square brackets in (3) do not appear in any other row, the above NLS problem can be split into ℓ separate NLS problems, one for each output of the system. Actually, the Jacobian J of the problem (3) is a block diagonal matrix,

$$J = \begin{bmatrix} J_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & J_\ell \end{bmatrix},$$

where J_s are full matrices of equal size, which can be computed analytically, if the basis functions were chosen (for instance, $\phi(\cdot) = \tanh(\cdot)$).

This step is referred to as the (nonlinear) initialization step.

Step 3: Estimating all parameters

The estimated system matrices from Step 1 and the estimated parameters in the vector θ from Step 2 are used as initial estimates to compute the parameters in a fully parameterized Wiener system with a fixed order of the state vector. To reduce the number of parameters for the linear part, the so-called *output normal form* parameterization [8], is used; that is, the pair (A, C) is transformed to satisfy

$$A^T A + C^T C = I_n, \quad (4)$$

where $I_n \in \mathbb{R}^{n \times n}$ denotes the identity matrix of order n . The condition (4) requires that the system matrix A be asymptotically stable.

Defining

$$\begin{bmatrix} C \\ A \end{bmatrix} = T_1 T_2 \cdots T_n \begin{bmatrix} 0 \\ I_n \end{bmatrix}, \quad (5)$$

where, for $k = 1, 2, \dots, n$,

$$\begin{aligned} T_k &= \begin{bmatrix} I_{k-1} & & \\ & U_k & \\ & & I_{n-k} \end{bmatrix} \in \mathbb{R}^{(n+\ell) \times (n+\ell)}, \quad U_k = \begin{bmatrix} -s_k & S_k \\ r_k & s_k^T \end{bmatrix}, \\ t_k &= s_k^T s_k, \quad r_k = \sqrt{1 - t_k}, \quad S_k = I_\ell - \frac{1 - r_k}{t_k} s_k s_k^T, \end{aligned}$$

then the pair (A, C) satisfies the condition (4). Each unitary matrix T_k is completely defined by a matrix U_k which is parameterized by the entries of the vector $s_k \in \mathbb{R}^\ell$. Hence, the total number of parameters needed for the pair (A, C) equals $n\ell$. From the formulas above, it follows that the parameter vectors s_k must satisfy $\|s_k\| < 1$. To avoid solving a constrained NLS, a bijective mapping,

$$h : \mathbb{R}^\ell \mapsto U_1(0) \subset \mathbb{R}^\ell, \quad s_k = h(\hat{s}_k), \quad U_1(0) = \{s \mid \|s\| < 1\}, \quad (6)$$

is used. Then, it is possible to perform the optimization with respect to the unconstrained vectors $\hat{s}_k := h^{-1}(s_k)$. The mapping used in the codes, and its inverse, are

$$h(\hat{s}_k) := \frac{2}{\pi} \cdot \frac{\arctan(\|\hat{s}_k\|)}{\|\hat{s}_k\|} \hat{s}_k, \quad h^{-1}(s_k) = \frac{\tan(\|s_k\|\frac{\pi}{2})}{\|s_k\|} s_k. \quad (7)$$

If the vectors \hat{s}_k together with the entries of the matrix pair (B, D) , stored column-wise, are stacked in the vector θ_{on} , then the parameter estimation problem of a fully parameterized Wiener system can be stated as

$$\min_{\theta, \theta_{on}, x(1)} \sum_{k=1}^N \|y(k) - \hat{y}(k, \theta, \theta_{on}, x(1))\|^2, \quad (8)$$

where $\hat{y}(k, \theta, \theta_{on}, x(1))$ is the output of the Wiener model

$$\begin{aligned} \hat{x}(k+1) &= A(\theta_{on})\hat{x}(k) + B(\theta_{on})u(k), \quad \hat{x}(1) = x(1), \\ \hat{z}(k) &= C(\theta_{on})\hat{x}(k) + D(\theta_{on})u(k), \\ \hat{y}_s(k, \theta, \theta_{on}, x(1)) &= \hat{f}_s(\hat{z}(k)), \quad s = 1, \dots, \ell. \end{aligned}$$

The initial estimates used in this NLS problem are obtained by Steps 1 and 2. The Jacobian matrix for (8) has the structure

$$J = \begin{bmatrix} J_1 & 0 & \cdots & 0 & J_1^l \\ 0 & J_2 & \cdots & 0 & J_2^l \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & J_\ell & J_\ell^l \end{bmatrix} \in \mathbb{R}^{\ell N \times (\ell((\ell+2)\nu+1)+n(\ell+m+1)+\ell m)},$$

where the last block column, with full submatrices, corresponds to the linear part. This block column is computed using a forward-difference approximation. The Cholesky factor R of $J^T J$, or the R factor of a QR factorization of J have a similar structure. The implementation stores J and R in a compressed form, using only two block columns.

The numerical optimization

The NLS problems (3) and (8) are solved using Levenberg-Marquardt algorithm [6], [7], [9], which iteratively finds a local minimum of the nonlinear cost function. These problems have the general structure

$$\min_{\Theta} \|y - \hat{y}(\Theta)\|^2, \quad (9)$$

where Θ is the stacked vector of unknown parameters, y is a given vector (e.g., $\text{vec}([y(k)]_{k=1}^N)$) and $\hat{y}(\Theta)$ is a given function. Define $e(\Theta) = y - \hat{y}(\Theta)$, the *residual* vector, whose components are the *error functions*.

Denoting $\Theta(l)$ the value of the parameter vector in the l -th iteration of the Levenberg-Marquardt algorithm, then this algorithm computes

$$\Theta(l+1) = \Theta(l) + \Delta\Theta(l),$$

with $\Delta\Theta(l)$ the solution of the positive definite set of linear equations

$$(J^T(l)J(l) + \mu I) \Delta\Theta(l) = -J^T(l)e(\Theta(l)). \quad (10)$$

The regularization coefficient $\mu \in (0, \infty)$, called *Levenberg factor*, is essential for the convergence of the algorithm. If μ is too small, the algorithm may diverge. For big values of μ , the convergence is very slow. The algorithm tries to keep μ as small as possible. If the cost function decreases, the current step is accepted, and the ratio of the actual decrease compared to the predicted decrease is checked. Then μ is decreased if the ratio was acceptable and increased otherwise. If the cost function increases, the step is rejected and the calculations are repeated with an increased μ . The algorithm convergence close to a local minimizer is quadratic for analytically computed Jacobian matrices, and linear for numerically computed Jacobian matrices [4].

Two versions of the Levenberg-Marquardt algorithm have been implemented: a MINPACK-like implementation, and a standard implementation. The first scheme uses a structure-exploiting QR decomposition with block-column pivoting of the matrix $J(l)$, while the second scheme uses either a Cholesky decomposition of the matrix $[J^T(l)J(l) + \mu I]$, built using its structure, or a CG algorithm [3] to solve (10).

3. Software Components

The current version of the SLICOT Wiener system identification toolbox consists of over 30 driver and computational routines, six MATLAB interfaces (MEX-files), and associated M-files. Many details are given in [10].

The MEX-files correspond to the Fortran drivers and other essential routines, and are listed in Table 1. The M-files call the MEX-files and are included for user's convenience. The same interfaces could also be used in a Scilab environment [2].

Table 1. Wiener system identification: MEX-file interfaces to MATLAB/Scilab.

<i>MEX-file</i>	<i>Function</i>
wident	Computes a discrete-time model of a Wiener system using a neural network approach and a MINPACK-like Levenberg-Marquardt algorithm.
widentc	Computes a discrete-time model of a Wiener system using a neural network approach and a Levenberg-Marquardt algorithm, based on either a Cholesky, or a conjugate gradients algorithm for solving (10).
Wiener	Computes the output of a Wiener system.
ldsim	Computes the output response of a linear discrete-time system (much faster than the MATLAB function lsim).
onf2ss	Transforms a linear discrete-time system given in the output normal form to a state-space representation.
ss2onf	Transforms a state-space representation of a linear discrete-time system into the output normal form.

The main MEX-files are `wident` and `widentc`, but the remaining MEX-files offer additional flexibility. Their calling sequences are

```
[xopt[,perf,nf,rcnd]]=wident(job,u,y,nn[,s,n,x,iter,nprint,
                                tol,seed,printw,ldwork]);
[xopt[,perf,nf,rcnd]]=widentc(job,u,y,nn[,s,n,x,alg,stor,
                                iter,nprint,tol,seed,printw,
                                ldwork])
```

where the parameters put inside the brackets are optional, and have default values. The parameters `u` and `y` denote the input and output trajectories, each row containing all input and output values, respectively, measured at the same time moment. The parameter `job` specifies which part of the Wiener parameterization must be initialized: the linear part only (`job = 1`); the static nonlinearity only (`job = 2`); both linear and nonlinear parts (`job = 3`); nothing, `x` already contains an initial approximation for Θ (`job = 4`). The parameters `nn`, `s`, and `n` denote the number of neurons, the number of block rows in the input and

output block-Hankel matrices processed for estimating the linear part (if `job` is 1 or 3), and the order of the linear part (or an option on how to compute it), respectively. The value of `n` should be given if `job` is 2 or 4. If `job` ≠ 3, the parameter `x` must contain the part of the vector of initial parameters specified by `job`. The parameters `iter` and `tol` are vectors with two elements, giving the maximal number of iterations and the tolerances for the initialization step, and the whole optimization. The parameter `nprint` specifies the frequency of printing the error norm in the iterative process. If `job` is 2 or 3, the parameter `seed` is a vector of length 4 containing the random number generator seed used to initialize the parameters of the static nonlinearity. Using `seed` enables to obtain reproducible results. The parameter `printw` is a switch for printing the warning messages. The parameter `ldwork` allows the user to specify other sizes than the default ones for working arrays. The parameters `alg` and `stor` specify the algorithm for solving (10) (Cholesky or CG), and how the matrix $J^T(l)J(l)$ is stored for Cholesky algorithm (full or packed), respectively.

The parameter `xopt` returns the optimized values of the parameters describing the Wiener system. The optional output parameters `perf`, `nf`, and `rcnd` contain: various performance results, e.g., the maximum residual error norm, the total number of iterations (and CG iterations, if any) performed, the final Levenberg factor; the (total) number of function and Jacobian evaluations; and the reciprocal condition number estimates (if `job` is 1 or 3) for estimating the linear part, respectively.

There are four computational M-files which call some of the MEX-files. Their calling sequences are listed below:

```
y      = NNout(nn,l,wb,u);
[y(,x)] = dsim(sys,u(,x0));
[sys,x0] = o2s(n,m,l,theta(,apply));
theta   = s2o(sys,x0(,apply));
```

`NNout` computes the output of a set of neural networks, and the remaining M-files correspond to the last three MEX-files in Table 1. The argument `wb` contains the weights and biases of the neural network. The parameters `x0` and `x` are the initial state (default `x0 = 0`) and the final state of the system, respectively. The parameter `sys` is a discrete-time `ss` MATLAB system object, consisting in a state-space realization $\text{sys} = (A, B, C, D)$. It may alternately be replaced by the matrix tuple. The parameter `theta` denotes the parameters of the linear part, in the output normal form parameterization, and `apply` specifies if the bijective mapping (7) should be used or not (default: the mapping is not used). The other arguments have already been defined before.

4. Numerical Results

The first example is also included in the compressed Wiener system identification toolbox file, `Wident_mex.zip`, available from the SLICOT ftp site `ftp://wgs.esat.kuleuven.ac.be/`, directory

```
pub/WGS/SLICOT/MatlabTools/Windows/SLToolboxes/
```

This example has 3 inputs, 2 outputs, and $t = 5000$ input and output data samples. The first 1000 samples were used to estimate the Wiener system, but all samples serve for validating the identified system. The linear system order and the number of neurons in the hidden layer were chosen as $n = 4$ and $nn = 12$. The corresponding optimization problem has 1000 error functions and 128 unknown variables. The MEX-file `wident` solved the problem in less than 30 seconds on a 500 MHz PC with 128 Mb memory, for tolerances set to 0.0001. It required 61 iterations for the initialization step, and 12 iterations for the whole optimization. The total number of function and Jacobian evaluations was 431, and 70, respectively. The Euclidean norm of the error was 3.5299 for the linear model, but 1.2914 for the Wiener model. The MEX-file `widentc`, option CG, solved this example in 324 seconds, and the error norm was 1.3066.

Many numerical tests have been performed using the DAISY identification collection (<http://www.esat.kuleuven.ac.be/sista/daisy>). The results enable to compare the implemented algorithms and their options. All algorithms have been initialized with the same seed for the random number generator. Also, the same tolerances (usually, 0.0001) have been used in all compared computations. The most efficient algorithm is problem dependent, but the MINPACK-like approach should be preferred from a numerical point of view. It was almost always the most accurate, and often the most efficient algorithm in our tests.

Some typical results are described below. Table 2 summarizes comparative performance results when using SLICOT MEX-files `wident` and `widentc` on a set of applications defined in the first three columns of the table. The same numbering scheme for applications as in [12] has been used. The estimation set consisted of the first half of each data set, nn was taken as 12, and the data were detrended. The original DAISY Application 16 has 28 outputs, but the first 7 outputs only have been considered, since the calculations are very time consuming. Even with this size reduction, the number of error functions is 4261, and the number of unknown parameters is 977.

As illustrations, the mean values of errors for linear and Wiener identification (computed for a moving window of 40 samples) are plotted in Figure 1 and Figure 2 for Applications 13 and 16, respectively. The improvement when using a nonlinear model is clearly visible.

Table 2. Comparative results for DAISY applications using `wident` and `widentc`.

Problem size			Execution time (sec.)			Sum of squares		
#	t	n, m, ℓ	<code>wident</code>	<code>Chol.</code>	<code>CG</code>	<code>wident</code>	<code>Chol.</code>	<code>CG</code>
2	1247	5,3,6	479.72	1158.98	10505.32	1.50e+1	8.39e+0	8.42e+0
5	2001	6,2,1	25.26	32.79	44.27	1.30e+1	1.30e+1	1.31e+1
6	867	10,3,3	360.36	40.43	108.86	5.46e+0	8.59e+0	8.81e+0
7	4000	6,1,1	8.73	17.25	46.57	1.30e+1	1.28e+1	1.28e+1
9	7500	5,1,2	1338.92	211.58	662.45	6.95e+0	1.34e+1	1.30e+1
10	9600	9,4,4	1471.29	588.85	3606.03	6.57e+2	2.10e+3	1.91e+3
11	1000	1,1,1	4.18	0.77	0.83	8.21e-1	9.53e-1	9.53e-1
12	1000	4,1,1	4.28	0.72	0.77	1.75e+0	1.79e+0	1.79e+0
14	1024	6,1,1	4.89	4.23	7.30	3.47e+0	3.39e+0	3.38e+0
15	1024	4,1,1	2.58	9.23	11.54	1.78e-1	1.71e-1	1.71e-1
16	8523	20,2,7	7956.51	3481.84	98595.72	1.55e+2	1.79e+2	1.55e+2
19	1680	3,2,1	6.04	3.63	6.76	3.87e+0	3.82e+0	3.82e+0
20	801	7,1,1	4.89	36.20	86.67	2.06e+1	1.04e+2	6.20e+1
21	99999	2,0,1	25.43	18.29	14.28	75532.0	75533.0	75533.0

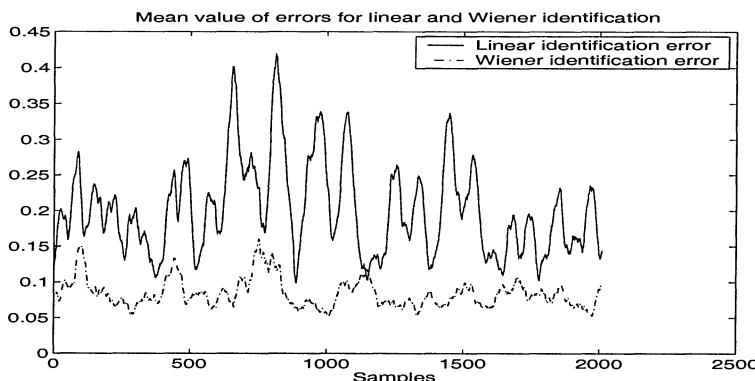


Figure 1. The mean values of errors for linear and Wiener identification, Application 13; all data samples used for estimation.

Finally, Table 3 summarizes comparative performance results when using SLICOT MEX-files `wident` and `widentc` for Application 16 for the first 7 outputs only. The numbers appearing in parentheses are the performance results corresponding to the initialization step.

5. Summary

Algorithmic, implementation and numerical details concerning nonlinear, multivariable Wiener systems identification have been investigated.

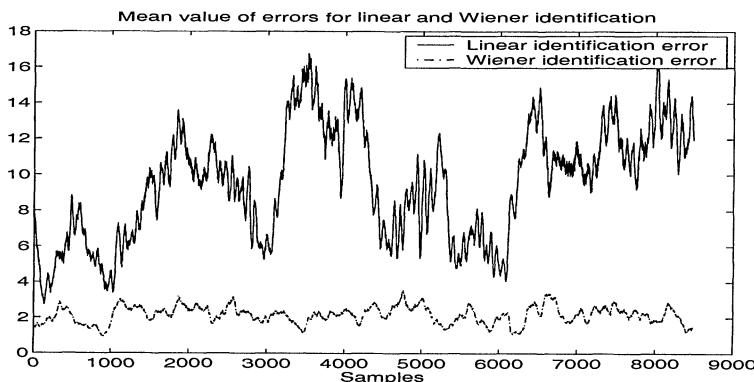


Figure 2. The mean values of errors for linear and Wiener identification, Application 16 (the first 7 outputs only); the first half of the data set used for estimation.

Table 3. Comparative results for Application 16 using `wident` and `widentc`.

Performance parameter	<code>wident</code>	<code>widentc, Cholesky</code>	<code>widentc, CG</code>
Execution time (sec.)	7956.51	3481.84	98595.72
Sum of squares	154.7	179.34	155.33
Number of iterations	31 (157)	19 (178)	64 (214)
Number of CG iterations	0 (0)	0 (0)	96206 (19769)
Total number of function evaluations	6658	4392	14155
Total number of Jacobian evaluations	184	197	278
Euclidean norm of the error for a Wiener model	225.08	249.03	225.7

A systematic three-step procedure for solving this problem has been used. Either a conjugate gradients algorithm or a direct, Cholesky-based algorithm is used for solving the linear systems of equations appearing in the computational process. Alternately, a MINPACK-like, specialized LAPACK-based Levenberg-Marquardt algorithm can be used, which proved to be very accurate and fast. The techniques are implemented in the new nonlinear system identification toolbox for the SLICOT Library. This toolbox includes interfaces (MEX-files and M-files) to the MATLAB and Scilab environments, which improve the user-friendliness of the collection. The results obtained show that the algorithms included in the toolbox are operational, and very fast.

References

- [1] Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., and Sorensen, D. (1999). *LAPACK Users' Guide: Third Edition*. Software · Environments · Tools. SIAM, Philadelphia.
- [2] Delebecque, F. and Steer, S. (1997). *Integrated Scientific Computing with Scilab*. Birkhäuser, Boston.
- [3] Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*. M. D. Johns Hopkins University Press, Baltimore, Maryland, third edition.
- [4] Kelley, C. T. (1999). *Iterative Methods for Optimization*. SIAM, Philadelphia, PA.
- [5] Mastronardi, N., Kressner, D., Sima, V., Van Dooren, P., and Van Huffel, S. (2001). A fast algorithm for subspace state-space system identification via exploitation of the displacement structure. *J. Comput. Appl. Math.*, 132(1):71–81.
- [6] Moré, J. J. (1978). The Levenberg-Marquardt algorithm: Implementation and theory. In Watson, G. A., editor, *Numerical Analysis*, volume 630 of *Lecture Notes in Mathematics*, pages 105–116. Springer-Verlag, Berlin, Heidelberg and New York.
- [7] Moré, J. J., Garbow, B. S., and Hillstrom, K. E. (1980). User's guide for MINPACK-1. Report ANL-80-74, Applied Math. Division, Argonne National Laboratory, Argonne, Illinois.
- [8] Peeters, R., Hanzon, B., and Olivi, M. (1999). Balanced realizations of discrete-time stable all-pass systems and the tangential Schur algorithm. In *Proceedings of the European Control Conference 31 August–3 September 1999, Karlsruhe, Germany*. Session CP-6, Discrete-time Systems.
- [9] Press, W. H., Teukolsky, S. A., Wetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes. The Art of Scientific Computing*. Cambridge University Press, New York, second edition.
- [10] Schneider, R., Riedel, A., Verdult, V., Verhaegen, M., and Sima, V. (2002). SLICOT system identification toolbox for nonlinear Wiener systems. SLICOT Working Note 2002-6, Katholieke Universiteit Leuven (ESAT/SISTA), Leuven, Belgium. Available from <ftp://wgs.esat.kuleuven.ac.be/pub/WGS/REPORTS/SLWN2002-6.ps.Z>.
- [11] Sima, V. (1999). Cholesky or QR factorization for data compression in subspace-based identification ? In *Proceedings of the Second NICONET Workshop on "Numerical Control Software: SLICOT, a Useful Tool in Industry", December 3, 1999, INRIA Rocquencourt, France*, pages 75–80.
- [12] Sima, V. and Van Huffel, S. (2001). Performance Investigation of SLICOT System Identification Toolbox, In *Proceedings of the European Control Conference, ECC 2001, 4–7 September, 2001, Seminário de Vilar, Porto, Portugal*, pages 3586–3591.
- [13] Verhaegen, M. (1993). Subspace model identification. Part 3: Analysis of the ordinary output-error state-space model identification algorithm. *Int. J. Control.*, 58(3):555–586.
- [14] Verhaegen, M. (1998). Identification of the temperature-product quality relationship in a multi-component distillation column. *Chemical Engineering Communications*, 163:111–132.
- [15] Westwick, D. and Verhaegen, M. (1996). Identifying MIMO Wiener systems using subspace model identification methods. *Signal Processing*, 52:235–258.

OPTIMIZATION OF DIFFERENTIAL SYSTEMS WITH HYSTERESIS

Jürgen Sprekels

Weierstrass Institute for Applied Analysis and Stochastics,
Mohrenstrasse 39, D – 10117 Berlin, Germany
sprekels@wias-berlin.de

Dan Tiba

Weierstrass Institute for Applied Analysis and Stochastics,
Mohrenstrasse 39, D – 10117 Berlin, Germany, and
Institute of Mathematics, Romanian Academy,
P. O. Box 1–764, RO–70700 Bucharest, Romania
tiba@wias-berlin.de and dtiba@imar.ro

Abstract We investigate general control problems governed by ordinary differential systems involving hysteresis operators. Our main hypotheses are of continuity type, and we discuss existence results, discretization methods, and approximation approaches.

Keywords: Peano-type existence theorem, continuous hysteresis operators, Clarke's generalized gradient

1. Introduction

Many engineering systems contain nonlinear functional dependencies of hysteresis type. We quote the recent monographs by Visintin [16] and by Brokate and Sprekels [6] for a mathematical treatment of the topic. Concerning the control of such systems basic references are the book of Brokate [2] and his articles [3], [4], [5], and the works of Smith [14], Banks, Smith and Wang [1].

We analyze a controlled ordinary differential system with hysteresis:

$$z' = f(t, z, y, u) \quad \text{in } [0, T], \quad (1.1)$$

$$z(0) = z_0 \in I\!\!R^N, \quad (1.2)$$

$$y(t) = W(S[z])(t), \quad S[z](t) = g(z(t)) \quad \text{in } [0, T], \quad (1.3)$$

$$u(t) \in U \quad \text{in } [0, T]. \quad (1.4)$$

Here $f : [0, T] \times \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}^N$ and $g : \mathbb{R}^N \rightarrow \mathbb{R}$ are continuous mappings, $U \subset \mathbb{R}^m$ is a closed bounded convex set, and $W : C[0, T] \rightarrow C[0, T]$ is a hysteresis operator, i.e. rate-independent and with the Volterra property, Sprekels and Brokate [6].

To the relations (1.1)–(1.4) the following cost functional is associated

$$J(u) = \int_0^T L(y(t), z(t), u(t)) dt, \quad (1.5)$$

with $L : \mathbb{R} \times \mathbb{R}^N \times U \rightarrow \mathbb{R}$ continuous in y, z , convex and lower semicontinuous with respect to u .

We allow g or S to be nonlinear, and many of the results that we shall establish will use just continuity and not require local Lipschitz assumptions on the data. Our investigation has the main motivation to provide a theoretically founded way towards the approximation and the numerical analysis of the control problem (1.1)–(1.5). In this respect, it seems that only the paper by Brokate [3] reports numerical experiments in a control problem with hysteresis. While in that work the optimality conditions are solved numerically, our approach uses a complete discretization of (1.1)–(1.5) and the computation of descent directions via the Clarke [8] generalized gradient. This allows the application of bundle-type algorithms, Strodiot and Nguyen [15], Lemaréchal [11].

In Section 2 we give the formulation of the control problem, and we establish the existence of optimal controls under general assumptions. Section 3 introduces the fully discretized optimization problem and studies existence and approximation questions. It is shown that the mapping control \mapsto state is Lipschitz under the given assumptions on f, g, W .

Section 4 uses an alternative formulation of the problem to analyze the variations and the directional derivatives. An algorithm and examples are also indicated.

2. Existence

We start with a Peano-type result for the Cauchy problem with hysteresis (1.1)–(1.3). We first omit the dependence on u , and we assume that $f : [0, T] \times \mathbb{R}^N \times \mathbb{R} \rightarrow \mathbb{R}^N$, $g : \mathbb{R}^N \rightarrow \mathbb{R}$, and $W : C[0, T] \rightarrow C[0, T]$ are continuous mappings and operators. Notice that W remains causal and rate-independent, i.e.:

$$v_1|_{[0,t]} = v_2|_{[0,t]} \Rightarrow W(v_1)(t) = W(v_2)(t), \quad t \in [0, T], \quad (2.1)$$

$$W(v \circ \varphi)(t) = W(v)(\varphi(t)), \quad \forall v \in C[0, T], \quad (2.2)$$

for any admissible time transformation $\varphi : [0, T] \rightarrow [0, T]$, continuous, nondecreasing, and onto.

Theorem 2.1 *Under the above assumptions, the initial value problem (1.1) – (1.3) has at least one global solution $z \in C^1([0, T]; \mathbb{R}^N)$ if the following sublinearity hypotheses are fulfilled:*

$$|W(w)|_{C[0,T]} \leq \alpha + \beta|w|_{C[0,T]}, \quad \forall w \in C[0,T], \quad (2.3)$$

$$|f(t, z, y)|_{\mathbb{R}^N} \leq \alpha + \beta|z|_{\mathbb{R}^N} + \gamma|y|, \quad (2.4)$$

$$|g(y)| \leq \alpha + \gamma|y|, \quad \alpha, \beta, \gamma \in \mathbb{R}_+. \quad (2.5)$$

Proof. For the positive numbers a, b and $c = \alpha + \beta q, q = \max\{|g(z)|; |z - z_0|_{\mathbb{R}^N} \leq b\}$, consider the set

$$\Delta = \{0 \leq t \leq a, |z - z_0|_{\mathbb{R}^N} \leq b, |y| \leq c\}.$$

Denote by $M = \max\{|f(t, z, y)|_{\mathbb{R}^N}; (t, z, y) \in \Delta\}$ and $\delta = \inf\left(a, \frac{b}{M}, \frac{c}{M}\right)$.

Clearly f is uniformly continuous in Δ , that is, $|f(t, z, y) - f(\tilde{t}, \tilde{z}, \tilde{y})|_{\mathbb{R}^N} < \varepsilon$ for any $\varepsilon > 0$ if $(t, z, y), (\tilde{t}, \tilde{z}, \tilde{y}) \in \Delta$ and $|t - \tilde{t}| < \eta(\varepsilon), |z - \tilde{z}|_{\mathbb{R}^N} < \eta(\varepsilon), |y - \tilde{y}| < \eta(\varepsilon)$. Denote by $h_\varepsilon = \inf\left(\eta(\varepsilon), \frac{\eta(\varepsilon)}{M}\right)$, and take the division $t_j = j h_\varepsilon, j \in \mathbb{N}$, of $[0, \delta]$.

We consider the polygonal functions (the Picard iterations with Euler polygonal lines):

$$\begin{aligned} \varphi_\varepsilon(t) &= \varphi_\varepsilon(t_j) + (t - t_j) f(t_j, \varphi_\varepsilon(t_j), y_\varepsilon(t_j)), \quad t_j < t \leq t_{j+1}, \\ \varphi_\varepsilon(0) &= z_0, \\ y_\varepsilon(t_j) &= W_f(S(\varphi_\varepsilon))(t_j). \end{aligned} \quad (2.6)$$

Due to (2.1), relation (2.5) makes sense. Recall that

$$W_f(S(\varphi_\varepsilon))(t_j) = \tilde{W}_f(S(z_0), S(\varphi_\varepsilon(t_1)), \dots, S(\varphi_\varepsilon(t_j))) \quad (2.7)$$

with $W_f : C[0, T] \rightarrow \mathbb{R}$ being the generating functional of W , Brokate and Sprekels [6]. Here, S is the set of all finite strings of real numbers and \tilde{W}_f the application induced on S by W_f (see Section 3).

Note that (2.2) plays an essential role in this construction. Then:

$$\begin{aligned} |\varphi_\varepsilon(t_{j+1}) - z_0|_{\mathbb{R}^N} &\leq |\varphi_\varepsilon(t_j) - z_0|_{\mathbb{R}^N} + h_\varepsilon |f(t_j, \varphi_\varepsilon(t_j), y_\varepsilon(t_j))| \\ &\leq |\varphi_\varepsilon(t_j) - z_0|_{\mathbb{R}^N} + M h_\varepsilon. \end{aligned} \quad (2.8)$$

Here, we argue by induction:

Assuming that $(t_j, \varphi_\varepsilon(t_j), y_\varepsilon(t_j)) \in \Delta$ and $|\varphi_\varepsilon(t_j) - z_0|_{\mathbb{R}^N} \leq j M h_\varepsilon$, relation (2.7) gives that $|\varphi_\varepsilon(t_{j+1}) - z_0| \leq (j+1) M h_\varepsilon \leq M \delta \leq b$.

By (2.3), (2.7), and (2.7), we have

$$|y_\varepsilon(t_j)| \leq \alpha + \beta \max_{0 \leq i \leq j} |S(\varphi_\varepsilon(t_i))|_{\mathbb{R}^N} \leq \alpha + \beta q = c. \quad (2.9)$$

Inequalities (2.7), (2.9) show that $(t_j, \varphi_\varepsilon(t_j), y_\varepsilon(t_j)) \in \Delta$ in all steps (2.5) such that $j h_\varepsilon \in [0, \delta]$. Then, $|f(t_j, \varphi_\varepsilon(t_j), y_\varepsilon(t_j))|_{\mathbb{R}^N} \leq M$, and (2.5) gives directly that

$$|\varphi_\varepsilon(t) - \varphi_\varepsilon(s)|_{\mathbb{R}^N} \leq M|t - s| \leq \eta(\varepsilon), \quad \forall \varepsilon > 0, \quad \forall t, s \in [0, \delta]. \quad (2.10)$$

Then $\varphi_\varepsilon \rightarrow \varphi$ uniformly in $[0, \delta]$, on a subsequence, and $\varphi \in C([0, \delta]; \mathbb{R}^N)$. If $\tilde{y}_\varepsilon(t) = W(S(\varphi_\varepsilon))(t)$, then $\tilde{y}_\varepsilon(t_j) = y_\varepsilon(t_j)$ and $\tilde{y}_\varepsilon \rightarrow y = W(S(\varphi))$ in $C[0, \delta]$ due to the continuity of S, W , and on the same subsequence. We also have that

$$\varphi'_\varepsilon(t) = f(t_j, \varphi_\varepsilon(t_j), y_\varepsilon(t_j)) \text{ if } t \in]t_j, t_{j+1}[, \quad (2.11)$$

while, by (2.10),

$$|f(t_j, \varphi_\varepsilon(t_j), y_\varepsilon(t_j)) - f(t, \varphi_\varepsilon(t), y_\varepsilon(t))|_{\mathbb{R}^N} \leq \varepsilon, \quad t \in [t_j, t_{j+1}] \quad (2.12)$$

by the uniform continuity of f in Δ .

For any $\lambda > 0$, again the uniform continuity of f in Δ gives that

$$|f(t, \varphi_\varepsilon(t), y_\varepsilon(t)) - f(t, \varphi_\varepsilon(t), \tilde{y}_\varepsilon(t))|_{\mathbb{R}^N} \leq \lambda, \quad t \in [t_j, t_{j+1}], \quad (2.13)$$

if $\varepsilon < \varepsilon(\lambda)$, due to the equicontinuity of the sequence $\{\tilde{y}_\varepsilon\}$ and to $\tilde{y}_\varepsilon(t_j) = y_\varepsilon(t_j)$.

Relations (2.11)–(2.13) allow to pass to the limit and to see that φ, y give a solution of (1.1)–(1.3) in $[0, \delta]$ and that $\varphi \in C^1[0, \delta]$.

Under assumptions (2.3)–(2.5), it is well-known that the local solution is in fact a global one, Brokate and Sprekels [6], p. 126. ■

Remark. By Theorem 2.1, uniqueness may be not true for the state system (1.1)–(1.3). The control problem (1.1)–(1.5) has to be understood as a minimization over pairs: to each control we associate all the possible states. This is well-known in the setting of optimal control theory of ODEs, Cesari [7], or for singular control problems for PDEs, Lions [12].

We assume that the continuity of $L(\cdot, \cdot, u) : \mathbb{R} \times \mathbb{R}^N \rightarrow \mathbb{R}_+$ is uniform with respect to $u \in U$. The mapping f is affine with respect to u , i.e.

$$f(t, z, y, u) = f_1(t, z, y) + f_2(t, z, y)u \quad (2.14)$$

where $f_1 : [0, T] \times \mathbb{R}^N \times \mathbb{R} \rightarrow \mathbb{R}^N$, $f_2 : [0, T] \times \mathbb{R}^N \times \mathbb{R} \rightarrow \mathbb{R}^{m \times N}$ are continuous sublinear mappings (like in (2.4)).

As U is bounded, convex and closed, the admissible controls are in $L^\infty([0, T]; \mathbb{R}^m)$, and the mapping (2.14) will not satisfy the continuity requirements of Theorem 2.1 (with respect to t , via u). However, the argument from its proof can be repeated when u is continuous, and a simple approximation argument may be used for $u \in L^\infty([0, T]; \mathbb{R}^m)$. The state $z \in W^{1,\infty}([0, T]; \mathbb{R}^N)$ in this case.

Theorem 2.2 *Under the above assumptions, the optimal control problem (1.1) – (1.5) has at least one optimal triplet*

$$[u^*, z^*, y^*] \in L^\infty([0, T]; U) \times W^{1,\infty}([0, T]; \mathbb{R}^N) \times C[0, T].$$

Proof. Let $[u_n, z_n, y_n]$ be a minimizing sequence. Then $\{u_n\}$ is bounded in $L^\infty([0, T]; \mathbb{R}^m)$. By (2.4), (2.14) and (1.1), (1.2), we have

$$\begin{aligned} |z_n(\hat{t})|_{\mathbb{R}^N} &\leq |z_0|_{\mathbb{R}^N} + \int_0^{\hat{t}} |f(t, z_n(t), y_n(t), u_n(t))|_{\mathbb{R}^N} dt \\ &\leq |z_0|_{\mathbb{R}^N} + c \int_0^{\hat{t}} (\alpha + \beta|z_n(t)|_{\mathbb{R}^N} + \gamma|y_n(t)|) dt. \end{aligned}$$

We may consider the truncated functions \tilde{y}_n, \tilde{z}_n to the interval $[0, t]$, and we still have $\tilde{y}_n(t) = y_n(t) = W(S(\tilde{z}_n))(t)$. Then (2.3) applies, and we get:

$$\begin{aligned} |z_n(\hat{t})|_{\mathbb{R}^N} &\leq |z_0|_{\mathbb{R}^N} + c \int_0^{\hat{t}} \left(\alpha + \beta|z_n(t)|_{\mathbb{R}^N} + \alpha + \beta \sup_{s \in [0, t]} |z_n(s)|_{\mathbb{R}^N} \right) dt \\ &\leq |z_0|_{\mathbb{R}^N} + 2c\alpha\hat{t} + 2c\beta \int_0^{\hat{t}} \sup_{s \in [0, t]} |z_n(s)|_{\mathbb{R}^N} dt. \end{aligned}$$

Taking the supremum in both sides of the inequality above, we obtain

$$\sup_{t \in [0, \hat{t}]} |z_n(t)|_{\mathbb{R}^N} \leq |z_0|_{\mathbb{R}^N} + 2c\alpha T + 2c\beta \int_0^{\hat{t}} \sup_{s \in [0, t]} |z_n(s)|_{\mathbb{R}^N} dt,$$

and Gronwall's lemma shows that $\{z_n\}$ is bounded in $C([0, T]; \mathbb{R}^N)$. By (1.1), (1.2), (2.14) we see that $\{z_n\}$ is bounded in $W^{1,\infty}([0, T]; \mathbb{R}^N)$. And (2.3), (2.5), (1.3) give that $\{y_n\}$ is bounded in $C[0, T]$.

We denote $u_n \rightarrow \bar{u} \in U$ weakly* in $L^\infty([0, T]; \mathbb{R}^m)$ and $z_n \rightarrow \bar{z}$ in $C([0, T]; \mathbb{R}^N)$, on a subsequence. Then, the continuity of W, S gives that $y_n \rightarrow \bar{y}$ in $C[0, T]$ with $\bar{y} = W(S(\bar{z}))$.

Moreover, (2.4) and (2.14), via the Lebesgue theorem, ensure that

$$\int_0^{\hat{t}} f(t, z_n(t), y_n(t), u_n(t)) dt \rightarrow \int_0^{\hat{t}} f(t, \bar{z}(t), \bar{y}(t), \bar{u}(t)) dt$$

which shows that $[\bar{u}, \bar{z}, \bar{y}]$ satisfies (1.1)–(1.4), i.e. it is admissible for the control problem. Notice that

$$\left| \int_0^T L(z_n(t), y_n(t), u_n(t)) dt - \int_0^T L(\bar{z}(t), \bar{y}(t), u_n(t)) dt \right| \rightarrow 0$$

as the continuity of $L(\cdot, \cdot, u_n)$ is uniform with respect to u_n . Thus:

$$\liminf_{n \rightarrow \infty} \int_0^T L(\bar{z}(t), \bar{y}(t), u_n(t)) dt \geq \int_0^T L(\bar{z}(t), \bar{y}(t), \bar{u}(t)) dt$$

by the convexity and lower semicontinuity with respect to u of L and the weak* convergence of u_n .

The two last convergence properties show that $[\bar{u}, \bar{z}, \bar{y}]$ is optimal for the problem (1.1)–(1.5), and the proof is finished. ■

3. Discretization

Consider an equidistant partition $\{t_i\}_{i=0, \overline{k}}$ of $[0, T]$ of step-size $\Delta t > 0$. The discretized control problem is

$$\text{Min} \left\{ \sum_{j=1}^{k+1} L(y_j, z_j, u_j) \Delta t \right\} \quad (\mathbf{P}_k)$$

subject to $u_{i+1} \in U$, $i = \overline{0, k}$ and

$$z_{i+1} = z_i + \Delta t f(t_{i+1}, z_{i+1}, y_{i+1}, u_{i+1}), \quad (3.1)$$

$$y_{i+1} = \tilde{W}_f(S(z_0), S(z_1), \dots, S(z_{i+1})). \quad (3.2)$$

Here, $\tilde{W}_f(s) = W_f(\pi_A(s))$, $\forall s \in \mathcal{S}$, the set of all finite strings of real numbers, and z_0 is given. The application $\pi_A(s)$ is the piecewise linear interpolation operator with equidistant nodes in $[0, T]$ corresponding to the number of components of $s \in \mathcal{S}$, and W_f is the generating functional associated to W , Brokate and Sprekels [6]. The definition of \tilde{W}_f is essentially based on (2.2). We have the following result.

Proposition 3.1 *If f, g, \tilde{W}_f are continuous in their arguments and satisfy (2.3)–(2.5), then for every $\{u_{i+1}\}_{0, \overline{k}} \in U^{k+1}$, the equations (3.1), (3.2) have at least one solution $\{z_{i+1}\}_{0, \overline{k}} \in \mathbb{R}^{N(k+1)}$, $\{y_{i+1}\}_{0, \overline{k}} \in \mathbb{R}^{k+1}$.*

Proof. The argument is iterative, for every i . Assuming the solution defined at level i is known, then $z_{i+1} \in \mathbb{R}^N$ and $y_{i+1} \in \mathbb{R}$ are obtained as a fixed point of the continuous application (in finite dimensional spaces) defined by the right-hand side of (3.1).

If B_{i+1} is a “big” closed ball around 0, in \mathbb{R}^N , containing z_0, z_1, \dots, z_i in the interior of $\frac{1}{2}B_{i+1}$, then relations (2.3)–(2.5) show that $z_i + \Delta t f(t_{i+1}, z, y, u_{i+1}) \in B_{i+1}$ for Δt “small” if $z \in B_{i+1}$ and $y = \tilde{W}_f(S(z_0), S(z_1), \dots, S(z_i), S(z))$. Then, Brouwer’s fixed point theorem, Kelley [10], provides at least one solution z_{i+1} of (3.1), and $y_{i+1} = \tilde{W}_f(S(z_0), S(z_1), \dots, S(z_i), S(z_{i+1}))$.

The important remark in this argument is that, for $z \in B_{i+1}$, then $|z|_{\mathbb{R}^N} \leq r_{i+1}$ (the radius), and (2.3)–(2.5) generate a constant $\Gamma > 0$, independent of i , such that $|f(t_{i+1}, z, y, u_{i+1})|_{\mathbb{R}^N} \leq \Gamma r_{i+1}$. Then $\Delta t = \frac{T}{k} < (2\Gamma)^{-1}$ will be a satisfactory choice of Δt in the above argument, which is also independent of i . ■

Corollary 3.2 *Under the above assumptions, if L is convex and lower semicontinuous in u and continuous in y, z uniformly with respect to u , then the discrete control problem (P_k) has at least one optimal n -tuple $[(u_j^k), (z_j^k), (y_j^k)]_{j=1, k+1}$ in $U^{k+1} \times (\mathbb{R}^N)^{k+1} \times \mathbb{R}^{k+1}$.*

The argument is similar to that used in the proof of Theorem 2.2, and we omit it.

Obviously $\{u_i^k\}$ are bounded in \mathbb{R}^m for any k and for $i = \overline{1, k+1}$. We examine the boundedness properties of $\{z_i^k\}$, $\{y_i^k\}$.

By (2.4), (3.1), and the boundedness of U , we get

$$|z_{i+1}^k|_{\mathbb{R}^N} \leq |z_i^k|_{\mathbb{R}^N} + C \Delta t \left(1 + |z_{i+1}^k|_{\mathbb{R}^N} + |y_{i+1}^k| \right). \quad (3.3)$$

We also have, by (3.2), (2.3), (2.5), that

$$|y_{i+1}^k| \leq C \left(1 + \max_{0 \leq j \leq i+1} |z_j^k|_{\mathbb{R}^N} \right). \quad (3.4)$$

By (3.3), (3.4), taking the maximum with respect to the indices, we see

$$\max_{0 \leq j \leq i+1} |z_j^k|_{\mathbb{R}^N} \leq \max_{0 \leq j \leq i} |z_j^k|_{\mathbb{R}^N} + C \Delta t \left(1 + 2 \max_{0 \leq j \leq i+1} |z_j^k|_{\mathbb{R}^N} \right). \quad (3.5)$$

Here C is an “absolute” constant depending just on α, β, γ from (2.3)–(2.5) and on the bound of $\{u_i^k\}$ in \mathbb{R}^m .

Summing (2.5) with respect to i , we can infer that

$$\max_{0 \leq j \leq i+1} |z_j^k|_{\mathbb{R}^N} \leq \sum_{l=0}^i C \Delta t \left(1 + 2 \max_{0 \leq j \leq l+1} |z_j^k|_{\mathbb{R}^N} \right). \quad (3.6)$$

If Δt is “small”, the discrete Gronwall inequality shows that $\{z_i^k\}$ are bounded in \mathbb{R}^N with respect to k and to $i = \overline{0, k+1}$. Inequality (3.4) gives the same for $\{y_i^k\}$ in \mathbb{R} .

Let us now construct the φ_k as in the proof of Theorem 2.1:

$$\varphi'_k(t) = f(t_{i+1}, \varphi_k(t_{i+1}), y_k(t_{i+1}), u_k(t_{i+1})), \quad t \in]t_{i+1}, t_{i+2}]. \quad (3.7)$$

The mapping $y_k(t) = W(S(\varphi_k))(t)$ and clearly $\{\varphi_k\}$ is bounded in $W^{1,\infty}([0, T]; \mathbb{R}^N)$, $\{y_k\}$ is bounded in $C[0, T]$, and u_k (piecewise constant interpolation of u_i^k) is bounded in $L^\infty([0, T]; \mathbb{R}^m)$.

We thus have $\varphi_k \rightarrow \hat{\varphi}$ uniformly in $C([0, T]; \mathbb{R}^N)$, on a subsequence. By the continuity of S and of W , we get $y_k \rightarrow \hat{y} = W(S \hat{\varphi})$ in $C[0, T]$. We also may assume $u_k \rightarrow \hat{u}$, on the same subsequence, weakly* in $L^\infty([0, T]; \mathbb{R}^m)$. Now, compute the difference

$$\begin{aligned} D &= f(t, \varphi_k(t), u_k(t)) - f(t_{i+1}, \varphi_k(t_{i+1}), y_k(t_{i+1}), u_k(t_{i+1})), \\ t &\in]t_{i+1}, t_{i+2}], \end{aligned}$$

and take into account that $u_k(t) \equiv u_k(t_{i+1})$ in this interval, and (2.14). The uniform continuity of f_1, f_2 and the above uniform convergences show that $|D| \leq \varepsilon$ if k is big enough. One can pass to the limit to see that $\hat{\varphi}, \hat{y}, \hat{u}$ is an admissible pair for the original control problem, i.e. it satisfies (1.1)–(1.4). By comparing with the cost obtained in (\mathbf{P}_k) via the discretization of u^* (provided by Theorem 2.2) and by passing to the limit, we also have:

Theorem 3.3 *The triplet $[\hat{u}, \hat{\varphi}, \hat{y}]$ is optimal for the problem (1.1) – (1.5).*

The next two statements concern Lipschitz properties of the mappings that we are using. We omit the proofs which are rather direct.

Lemma 3.4 *If $W : C[0, T] \rightarrow C[0, T]$ is Lipschitz of rank $C > 0$ and $s_1 = (v_0, v_1, \dots, v_l), s_2 = (w_0, w_1, \dots, w_l) \in \mathcal{S}$ have the same number of components, then*

$$|\tilde{W}_f(v_0, v_1, \dots, v_l) - \tilde{W}_f(w_0, w_1, \dots, w_l)| \leq C \max_{0 \leq j \leq l} |v_j - w_j|. \quad (3.8)$$

Remark. Under regularity/Lipschitz assumptions on f, g, L , Lemma 3.4 shows that the functional dependence from $\{u_i\} \in U^{k+1}$ to the cost is a Lipschitzian dependence. Thus, the Clarke [8] generalized gradient may be used to write the optimality conditions for (\mathbf{P}_k) and to devise descent algorithms. More about this will be said below.

Proposition 3.5 *Assume that f, g are real Lipschitz mappings and that W is a Lipschitz operator in $C[0, T]$. Then the correspondence $\{u_i\} \mapsto \{z_i\}$ defined by (3.1), (3.2) is Lipschitz from $(\mathbb{R}^m)^{k+1}$ to $(\mathbb{R}^N)^{k+1}$.*

Remark. If the differentiable mapping $\tilde{S} : (\mathbb{R}^N)^{i+1} \rightarrow \mathbb{R}^{i+1}$, $\tilde{S}(z_0, z_1, \dots, z_{i+1}) = (S(z_0), S(z_1), \dots, S(z_{i+1}))$ has a surjective Jacobian, then the chain rule is valid for the Clarke generalized gradient “ ∂ ” of the composed mapping $\tilde{W}_f(\tilde{S})$, Clarke et al. [9], Theorem 3.2:

$$\partial(\tilde{W}_f \circ \tilde{S})(\cdot) = [\tilde{S}'(\cdot)]^* \partial \tilde{W}_f(\tilde{S}(\cdot)), \quad (3.9)$$

and one can write the first-order optimality conditions for (\mathbf{P}_k) .

4. Approximation

In this section, in order to fix ideas, we shall assume that the mapping L is quadratic and independent of y . We shall perform a further approximation of the problem (\mathbf{P}_k) by the penalization of (3.2) into the cost. We do not regularize the hysteresis operator W , as in Brokate [2]. Roughly speaking, we shall interpret y as a supplementary/artificial control, and (3.2) as a mixed control-state constraint.

Since in the theory of hysteresis operators, Brokate and Sprekels [6], the piecewise monotonicity of mappings plays an important role, our penalization method uses just the positive part function, $(\cdot)_+$, which is monotone:

$$\begin{aligned} \text{Min} \left\{ \frac{1}{2} \sum_{j=1}^{k+1} |z_j - z_d(t_j)|_{\mathbb{R}^N}^2 (t_{j+1} - t_j) + \frac{1}{2} \sum_{j=1}^{k+1} |u_j|_{\mathbb{R}^m}^2 (t_{j+1} - t_j) \right. \\ \left. + \frac{1}{\varepsilon} \max_{j=1, k+1} \left[\left(y_j - \tilde{W}_f(S(z_0), \dots, S(z_j)) \right)_+ ; \right. \right. \\ \left. \left. \left(\tilde{W}_f(S(z_0), \dots, S(z_j)) - y_j \right)_+ \right] \right\} \end{aligned} \quad (4.1)$$

subject to $u_j \in U$, $y_j \in \mathbb{R}$, (3.1), and with z_0 given as an initial condition.

The approximation properties of the penalized problem (4.1) with respect to (\mathbf{P}_k) , when $\varepsilon \rightarrow 0$, are standard, and we do not discuss this here.

Moreover, under usual differentiability and Lipschitz assumptions on f in (3.1) the mapping $\{y_i\} \in \mathbb{R}^{k+1}$, $\{u_i\} \in (\mathbb{R}^m)^{k+1} \mapsto \{z_i\} \in (\mathbb{R}^N)^{k+1}$ is differentiable and Lipschitz for the corresponding finite dimensional norms.

Consequently, we may view (4.1) as the minimization of a Lipschitzian real mapping (\tilde{W}_f is just Lipschitz), depending on $\{u_i\}$ and $\{y_i\}$, and under the constraint $u_i \in U$, convex, closed, bounded subset in \mathbb{R}^m .

We have already seen that the Clarke generalized gradient of $\tilde{W}_f(\tilde{S}(\cdot))$ may be computed via the chain rule (3.9). This may be done directly with respect to $\{u_i\}$ and $\{y_i\}$, since the dependence (3.1) of $\{z_i\}$ on these variables, denoted as $\mathcal{A} : \mathbb{R}^{k+1} \times (\mathbb{R}^m)^{k+1} \rightarrow (\mathbb{R}^N)^{k+1}$, may be assumed C^1 , and having a surjective Jacobian.

Let us denote shortly by $\mathcal{C} = \tilde{W}_f \circ \tilde{S} \circ \mathcal{A} : \mathbb{R}^{k+1} \times (\mathbb{R}^m)^k \rightarrow \mathbb{R}$, the superposition Lipschitzian mapping. It is to be noticed that the composition of \mathcal{C} with $(\cdot)_+$, appearing in (4.1), does not fulfil the assumptions of the chain rules indicated in Clarke et al. [9], Ch. 2.4. In particular, the mapping \mathcal{C} is not regular, in general (i.e. the generalized Clarke directional derivative may not coincide with the usual directional derivative). However, the mapping $(\cdot)_+$ is regular (since it is convex), has positive gradients and, clearly, a very simple structure. A direct computation may be used to establish the following result.

Proposition 4.1 *If $y_i - \mathcal{C}(\{u_i\}, \{y_i\}) > 0$, and if $[\{v_i\}, \{x_i\}]$ is some variation of $[\{u_i\}, \{y_i\}]$, respectively, then*

$$\begin{aligned} & \limsup_{\substack{\{\{\tilde{u}_i\}, \{\tilde{y}_i\} \rightarrow \{\{u_i\}, \{y_i\}\} \\ \lambda \downarrow 0}} \frac{(\tilde{y}_i + \lambda x_i - \mathcal{C}([\{\tilde{u}_i\}, \{\tilde{y}_i\}] + \lambda [\{v_i\}, \{x_i\}]))_+ - (\tilde{y}_i - \mathcal{C}([\{\tilde{u}_i\}, \{\tilde{y}_i\}]))_+}{\lambda} \\ &= \limsup_{\substack{\{\{\tilde{u}_i\}, \{\tilde{y}_i\} \rightarrow \{\{u_i\}, \{y_i\}\} \\ \lambda \downarrow 0}} \frac{\tilde{y}_i + \lambda x_i - \mathcal{C}([\{\tilde{u}_i\}, \{\tilde{y}_i\}] + \lambda [\{v_i\}, \{x_i\}]) - \tilde{y}_i + \mathcal{C}([\{\tilde{u}_i\}, \{\tilde{y}_i\}])}{\lambda}. \end{aligned}$$

If $y_i - \mathcal{C}(\{u_i\}, \{y_i\}) < 0$, the above \limsup is null.

Remark. By Clarke et al. [9], p. 79, if \tilde{S} and \mathcal{A} are C^1 with surjective Jacobians, then the last \limsup in Proposition 4.1 coincides with

$$\limsup_{\substack{w \rightarrow \tilde{S}(\mathcal{A}(\{u_i\}, \{y_i\})) \\ \lambda \downarrow 0}} \left[-\frac{\tilde{W}_f(w + \lambda(\tilde{S} \circ \mathcal{A})'(\{v_i\}, \{x_i\})) - \tilde{W}_f(w)}{\lambda} \right] + x_i. \quad (4.2)$$

Here, $(\tilde{S} \circ \mathcal{A})'$ denotes the Jacobi matrix of the composed mapping. Relation (4.2) and Proposition 4.1 indicate how to compute the Clarke generalized directional derivative of the cost functional (4.1). For the max-operation appearing in (4.1), one has to take the maximum of the \limsup computed as above.

The main example that we consider concerns the case when W is the so-called *play operator*. We introduce the real mapping (for some given $r > 0$)

$$f_r(v, w) = \max\{v - r, \min\{v + r, w\}\}, \quad \forall v, w \in \mathbb{R}. \quad (4.3)$$

Taking into account the discretized problems (3.2) or (4.1), we define directly the mapping \tilde{W}_f on \mathcal{S} . This can be done inductively, Brokate

and Sprekels [6], p. 39:

$$\tilde{W}_f(v_0) = f_r(v_0, w), \quad \forall v_0 \in \mathbb{R}, \quad (4.4)$$

$$\tilde{W}_f(v_0, \dots, v_i) = f_r(v_i, \tilde{W}_f(v_0, \dots, v_{i-1})), \quad \forall v_0, \dots, v_i \in \mathbb{R}. \quad (4.5)$$

Here $w \in \mathbb{R}$ is fixed (one can take $w = 0$) and has the significance of an initial (or memory) condition imposed on the operator \tilde{W}_f .

Proposition 4.2 *The nonlinear functional \tilde{W}_f is piecewise linear on \mathbb{R}^i , for any given i .*

This is an immediate consequence of (4.3)–(4.5). We underline that piecewise linear functionals are neither regular in the sense of Clarke [8], nor weakly semismooth in the sense of Mifflin [13].

However, Proposition 4.2 and (3.9) show that it is possible to compute numerically the Clarke generalized gradient $\partial\tilde{W}_f(\tilde{S} \circ \mathcal{A}(\cdot))$, in any point. The observation is that $\partial\tilde{W}_f(\cdot)$ is piecewise constant in \mathbb{R}^i , for any i and, consequently, a finite number of operations will suffice.

It is known that $d(\cdot) = \text{proj}_{\partial\tilde{W}_f(\tilde{S} \circ \mathcal{A}(\cdot))}\{0\}$ is a descent direction if it is nonzero, Clarke [8]. This may be tested by the computations indicated in Proposition 4.1. If it is zero, then a stationary point has been achieved. The following conceptual algorithm may be used:

Algorithm 4.3

1 Let $\{u_i\}, \{y_i\}$ be given.

2 Compute the generating vectors of $\partial J(\{u_i\}, \{y_i\})$.

3 Compute d .

4 If $d = 0$, then STOP.

5 If $d \neq 0$, then

$$[\{u_i\}, \{y_i\}] \rightarrow [\{u_i\}, \{y_i\}] - \rho d, \quad \rho > 0.$$

6 Compute $J([\{u_i\}, \{y_i\}] - \rho d)$.

7 GO TO 2.

Here J is the cost defined by (4.1). We note that in Step 5, a line search has to be performed. In general, there is no convergence ensured for Algorithm 4.3 as the weak semismoothness of J is not valid, Strodiot and Nguyen [15]. Therefore, in practice, a number of steps has to be prescribed or some numerical convergence tests have to be introduced.

References

- [1] H.T. Banks, R.C. Smith and Y. Wang. *Smart Material Structures: Modeling, Estimation and Control*. Masson, Paris, 1996.
- [2] M. Brokate. *Optimale Steuerung von gewöhnlichen Differentialgleichungen mit Nichtlinearitäten vom Hysteresis-Typ*. Peter-Lang-Verlag, Frankfurt am Main, 1987.
- [3] M. Brokate. Numerical solution of an optimal control problem with hysteresis. In *LN Control and Information Sciences 95*, pages 68–78, Berlin, 1987. Springer.
- [4] M. Brokate. Optimal control of ODE systems with hysteresis nonlinearities. In *ISNM 84*, pages 25–41, Basel, 1988. Birkhäuser.
- [5] M. Brokate. Optimal control of the semilinear wave equation with hysteresis. In *Free Boundary Problems: Theory and Applications* (K.H. Hoffmann and J. Sprekels, eds.), pages 451–458, Harlow, 1990. Longman.
- [6] M. Brokate and J. Sprekels. *Hysteresis and Phase Transitions*. Springer, New York, 1996.
- [7] L. Cesari. *Optimization — Theory and Applications*. Springer, Berlin, 1983.
- [8] F.H. Clarke. *Optimization and Nonsmooth Analysis*. Wiley Interscience, New York, 1983.
- [9] F.H. Clarke, Yu.S. Ledyaev, R.J. Stern and P.R. Wolenski. *Nonsmooth Analysis and Control Theory*. Springer Graduate Texts in Mathematics 178, New York, 1998.
- [10] J.L. Kelley. *General Topology*. Springer, New York–Berlin, 1975.
- [11] C. Lemaréchal. Bundle methods in nondifferentiable optimization. In *Nonsmooth Optimization* (C. Lemaréchal and R. Mifflin, eds.), pages 79–102, Oxford, 1978. Pergamon Press.
- [12] J.L. Lions. *Contrôle des systèmes distribués singuliers*. Dunod, Paris, 1983.
- [13] R. Mifflin. An algorithm for constrained optimization with semi-smooth functions. *Math. Oper. Res.*, 2:191–207, 1977.
- [14] R.C. Smith. Hysteresis modeling in magnetostrictive materials via Preisach operators. *J. Math. Systems Estim. Control*, 8(2):23 pp. (electronic), 1998.
- [15] J.J. Strodiot and V.H. Nguyen. On the numerical treatment of the inclusion $0 \in \partial f(x)$. In *Topics in Nonsmooth Mechanics* (J.J. Moreau, P.D. Panagiotopoulos, G. Strang, eds.), pages 267–294, Basel, 1988. Birkhäuser.
- [16] A. Visintin. *Differential Models of Hysteresis*. Springer, Berlin, 1994.

OPTIMAL CONTROL OF NON STATIONARY, THREE DIMENSIONAL MICROPOLAR FLOWS

Ruxandra Stavre

Institute of Mathematics, Romanian Academy

P.O. Box 1-764, RO-70700 Bucharest, Romania

rstavre@imar.ro

Abstract We study an optimal control problem associated with a nonstationary, three dimensional flow of a micropolar fluid. We consider a suitable formulation of the control problem which allows us to prove the existence of a solution of this problem and to obtain the necessary conditions of optimality.

Keywords: micropolar fluid, weak solution, strong solution, control problem.

1. Introduction

The flow of micropolar fluids is a problem of physical interest since animal blood, liquid crystals, certain polymeric fluids, etc may be represented by the mathematical model of these fluids. This model was introduced by Eringen in [1]. From the physical point of view, a micropolar fluid is characterized by the following property: fluid points contained in a small volume element, in addition to its usual rigid motion, can rotate about the centroid of the volume element in an average sense, the rotation being described by a skew-symmetric gyration tensor, ω .

In this paper we are concerned with the nonstationary, three dimensional incompressible motion of a micropolar fluid. As in the 3-D case of Navier-Stokes equations (see [2]), we define weak solutions and strong solutions of the system describing the micropolar flow, and it is known that in the class of weak solutions we cannot prove the uniqueness, while for strong solutions we obtain the uniqueness, but there is no an existence result.

The aim of this paper is to study an optimal control problem associated with the evolution system describing the flow of a micropolar fluid.

This type of problems, for two dimensional flows, has been studied by Stavre in [3], [4], [5]. For the 3-D case, the study is more complicated, since we cannot prove the existence of a strong solution. To overcome this difficulty, we consider a suitable formulation of the control problem (as in [6], [7]), which allows us to prove the existence of a solution of the control problem and to obtain the necessary conditions of optimality.

The paper is organized as follows: in Section 2 we introduce the system of coupled equations which describes the nonstationary, three dimensional flow of an incompressible micropolar fluid and its variational formulation. We discuss about weak and strong solutions of this system and about their existence and uniqueness. By proving a general result, we obtain the desired regularity for the unknowns of the problem. In the next section we formulate the control problem such that to every optimal control we can associate a strong solution. The existence of a solution of the considered control problem is investigated. The last section deals with the first order optimality conditions.

2. Analysis of the motion system

The nonstationary, incompressible, three dimensional motion of a micropolar fluid with non-homogeneous initial data is described by the following coupled system:

$$\begin{cases} \vec{v}' + (\vec{v} \cdot \nabla) \vec{v} - (\mu + \chi) \Delta \vec{v} + \nabla p - \chi \operatorname{curl} \vec{\omega} = \vec{f} & \text{in } \Omega_T, \\ j\vec{\omega}' + j(\vec{v} \cdot \nabla) \vec{\omega} - \gamma \Delta \vec{\omega} - (\alpha + \beta) \nabla(\operatorname{div} \vec{\omega}) + 2\chi \vec{\omega} - \chi \operatorname{curl} \vec{v} = \vec{g} & \text{in } \Omega_T, \\ \operatorname{div} \vec{v} = 0 & \text{in } \Omega_T, \\ \vec{v} = \vec{0}, \vec{\omega} = \vec{0} & \text{on } \partial\Omega \times (0, T), \\ \vec{v}(x, 0) = \vec{v}_0(x), \vec{\omega}(x, 0) = \vec{\omega}_0(x) & \text{in } \Omega, \end{cases} \quad (2.1)$$

where $\Omega \subset I\!\!R^3$ is an open, bounded, connected set, with $\partial\Omega$ of class C^2 , T a positive given constant and $\Omega_T = \Omega \times (0, T)$, $\chi, \mu, j, \alpha, \beta, \gamma$ are positive given constants associated with the properties of the material, \vec{f}, \vec{g} are the given external fields, $\vec{v}_0, \vec{\omega}_0$ are the initial data and $\vec{v}, \vec{\omega}, p$ are the unknown of the system: the velocity, the microrotation and the pressure of the micropolar fluid, respectively.

We shall need the following spaces (for their properties see, e.g. [2])

$$\begin{cases} V = \{\vec{u} \in (H_0^1(\Omega))^3 / \operatorname{div} \vec{u} = 0\}, \\ H = \{\vec{u} \in (L^2(\Omega))^3 / \operatorname{div} \vec{u} = 0, \vec{u} \cdot \vec{n}|_{\partial\Omega} = 0\}, \\ H^{2,1}(\Omega_T) = \{u \in L^2(\Omega_T) / u', \frac{\partial u}{\partial x_i}, \frac{\partial^2 u}{\partial x_i \partial x_j} \in L^2(\Omega_T); i, j = 1, 2\}, \end{cases}$$

The following notation will be used throughout the paper:

- (\cdot, \cdot) the scalar product , $|\cdot|$ the norm in $L^2(\Omega)$ or $(L^2(\Omega))^3$,
- $((\cdot, \cdot))_0$ the scalar product , $\|\cdot\|_0$ the norm in $H_0^1(\Omega)$ or $(H_0^1(\Omega))^3$,
- $\langle \cdot, \cdot \rangle_{X', X}$ the duality pairing between a space X and its dual X' ,
- $b(\vec{u}, \vec{v}) = (\vec{u} \cdot \nabla) \vec{v}, \quad \forall \vec{u}, \vec{v} \in (H_0^1(\Omega))^3$.

For $\vec{f} \in L^2(0, T; H)$, $\vec{g} \in L^2(0, T; (L^2(\Omega))^3)$, $\vec{v}_0 \in V$, $\vec{\omega}_0 \in (H_0^1(\Omega))^3$, the variational formulation of the problem (2.1) is given by

$$\left\{ \begin{array}{l} \langle \vec{v}'(t), \vec{z} \rangle_{V', V} + \langle b(\vec{v}(t), \vec{v}(t)), \vec{z} \rangle_{V', V} + (\mu + \chi)((\vec{v}(t), \vec{z}))_0 \\ -\chi(\operatorname{curl} \vec{\omega}(t), \vec{z}) = (\vec{f}(t), \vec{z}) \quad \forall \vec{z} \in V, \\ j \langle \vec{\omega}'(t), \vec{\eta} \rangle_{(H^{-1}(\Omega))^3, (H_0^1(\Omega))^3} + j \langle b(\vec{v}(t), \vec{\omega}(t)), \vec{\eta} \rangle_{(H^{-1}(\Omega))^3, (H_0^1(\Omega))^3} \\ + \gamma((\vec{\omega}(t), \vec{\eta}))_0 + (\alpha + \beta)(\operatorname{div} \vec{\omega}(t), \operatorname{div} \vec{\eta}) + 2\chi(\vec{\omega}(t), \vec{\eta}) \\ -\chi(\operatorname{curl} \vec{v}(t), \vec{\eta}) = (\vec{g}(t), \vec{\eta}) \quad \forall \vec{\eta} \in (H_0^1(\Omega))^3, \\ \vec{v}(0) = \vec{v}_0, \quad \vec{\omega}(0) = \vec{\omega}_0. \end{array} \right. \quad (2.2)$$

The next theorem gives the existence (without the uniqueness) of a weak solution and the uniqueness (without the existence) of a strong solution of the variational formulation (2.2).

Theorem 2.1. a) *There exists at least a pair $(\vec{v}, \vec{\omega})$ with the regularity $\vec{v} \in L^2(0, T; V) \cap L^\infty(0, T; H)$, $\vec{v}' \in L^{4/3}(0, T; V')$, $\vec{\omega} \in L^2(0, T; (H_0^1(\Omega))^3) \cap L^\infty(0, T; (L^2(\Omega))^3)$, $\vec{\omega}' \in L^{4/3}(0, T; (H^{-1}(\Omega))^3)$, satisfying (2.2) a. e. in $(0, T)$. Such a solution is called a weak one.*

b) *There exists at most a pair $(\vec{v}, \vec{\omega})$ which is a weak solution of (2.2) and satisfies $\vec{v} \in L^8(0, T; (L^4(\Omega))^3)$. This solution is called a strong solution of (2.2).*

Proof. The main steps in obtaining the results of point a) are similar to those for Navier-Stokes equations (see [2]). For proving the second assertion, we need further regularity of the function $\vec{\omega}$, (i. e. $\vec{\omega} \in L^8(0, T; (L^4(\Omega))^3)$). This regularity will be obtained in Corollary 2.3, proved below and will allow us to obtain the uniqueness of the strong solution.

In the sequel, we shall prove a general result, which will give the regularity of the solutions throughout the paper.

Theorem 2.2. *Let $\vec{f} \in L^2(0, T; H)$, $\vec{g} \in L^2(0, T; (L^2(\Omega))^3)$, $\vec{v}_0 \in V$, $\vec{\omega}_0 \in (H_0^1(\Omega))^3$, $\vec{u} \in L^8(0, T; (L^4(\Omega))^3)$, $\vec{y} \in L^\infty(0, T; V) \cap L^{3/2}(0, T; (H^2(\Omega))^3)$, and $\vec{\rho} \in (H^{2,1}(\Omega_T))^3 \cap C([0, T]; (H_0^1(\Omega))^3)$. Then there exists an unique pair $(\vec{v}, \vec{\omega}) \in (H^{2,1}(\Omega_T))^3 \cap C([0, T]; V) \times (H^{2,1}(\Omega_T))^3 \cap C([0, T]; (H_0^1(\Omega))^3)$ satisfying, together with a function $p \in L^2(0, T; H^1(\Omega))$, unique up to the addition of a function of t , the following system:*

$$\left\{ \begin{array}{l} \vec{v}' + b(\vec{v}, \vec{y}) + b(\vec{u}, \vec{v}) - (\mu + \chi) \Delta \vec{v} + \nabla p - \chi \operatorname{curl} \vec{\omega} = \vec{f} \text{ in } \Omega_T, \\ j\vec{\omega}' + jb(\vec{v}, \vec{\rho}) + jb(\vec{u}, \vec{\omega}) - \gamma \Delta \vec{\omega} - (\alpha + \beta) \nabla(\operatorname{div} \vec{\omega}) + \\ \quad + 2\chi \vec{\omega} - \chi \operatorname{curl} \vec{v} = \vec{g} \text{ in } \Omega_T, \\ \operatorname{div} \vec{v} = 0 \text{ in } \Omega_T, \\ \vec{v} = \vec{0}, \vec{\omega} = \vec{0} \text{ on } \partial\Omega \times (0, T), \\ \vec{v}(x, 0) = \vec{v}_0(x), \vec{\omega}(x, 0) = \vec{\omega}_0(x) \text{ in } \Omega, \end{array} \right. \quad (2.3)$$

Proof. For proving the existence and the regularity, we approximate the functions \vec{v} , $\vec{\omega}$ with

$$\vec{v}_m = \sum_{i=1}^m g_{im}(t) \vec{u}_i, \quad \vec{\omega}_m = \sum_{i=1}^m h_{im}(t) \vec{\varphi}_i, \quad (2.4)$$

where $\{\vec{u}_i\}_i$ is a base of V and $\{\vec{\varphi}_i\}_i$ a base of $(H_0^1(\Omega))^3$. We consider the variational formulation of (2.3) corresponding to $(\vec{v}_m, \vec{\omega}_m)$:

$$\left\{ \begin{array}{l} (\vec{v}'_m(t), \vec{u}_j) + (b(\vec{v}_m(t), \vec{y}(t)), \vec{u}_j) + (b(\vec{u}(t), \vec{v}_m(t)), \vec{u}_j) \\ + (\mu + \chi)((\vec{v}_m(t), \vec{u}_j))_0 - \chi(\operatorname{curl} \vec{\omega}_m(t), \vec{u}_j) = (\vec{f}(t), \vec{u}_j), \\ j(\vec{\omega}'_m(t), \vec{\varphi}_j) + j(b(\vec{v}_m(t), \vec{\rho}(t)), \vec{\varphi}_j) + j(b(\vec{u}(t), \vec{\omega}_m(t)), \vec{\varphi}_j) \\ + \gamma((\vec{\omega}_m(t), \vec{\varphi}_j))_0 + (\alpha + \beta)(\operatorname{div} \vec{\omega}_m(t), \operatorname{div} \vec{\varphi}_j) \\ + 2\chi(\vec{\omega}_m(t), \vec{\varphi}_j) - \chi(\operatorname{curl} \vec{v}_m(t), \vec{\varphi}_j) = (\vec{g}(t), \vec{\varphi}_j) \quad \forall j = 1, \dots, m, \\ \vec{v}_m(0) = \vec{v}_{0m}, \quad \vec{\omega}_m(0) = \vec{\omega}_{0m}, \end{array} \right. \quad (2.5)$$

with $\vec{v}_{0m} \rightarrow \vec{v}_0$ in V and $\vec{\omega}_{0m} \rightarrow \vec{\omega}_0$ in $(H_0^1(\Omega))^3$. If we introduce (2.4) into (2.5) we obtain a linear system of ordinary differential equations with the unique solution g_{im} , h_{im} , $i = 1, \dots, m$. For obtaining the existence of $(\vec{v}, \vec{\omega})$ we establish some *a priori* estimates. The estimates in $L^\infty(0, T; H) \times L^\infty(0, T; (L^2(\Omega))^3)$ and in $L^2(0, T; V) \times L^2(0, T; (H_0^1(\Omega))^3)$ are obtained in the classical way (see e.g. [2]). We establish next the estimates in $L^\infty(0, T; V) \times L^\infty(0, T; (H_0^1(\Omega))^3)$ and in $(L^2(0, T; (H^2(\Omega))^3))^2$. For this purpose, we define the linear operator $L: (H_0^1(\Omega))^3 \mapsto (H^{-1}(\Omega))^3$,

$$L\vec{\omega} = -\gamma \Delta \vec{\omega} - (\alpha + \beta) \nabla(\operatorname{div} \vec{\omega}), \quad \forall \vec{\omega} \in (H_0^1(\Omega))^3. \quad (2.6)$$

The linear operators $-\Delta$ and L being compact and self adjoint, they have an orthonormal sequence of eigenfunctions, which can be taken as a base in V and in $(H_0^1(\Omega))^3$, respectively. We take in (2.4)₁ the sequence of eigenfunctions of $-\Delta$ and in (2.4)₂ the sequence of eigenfunctions of L . Hence \vec{u}_i and $\vec{\varphi}_i$ satisfy:

$$\left\{ \begin{array}{l} -\Delta \vec{u}_i + \nabla p_i = \lambda_i \vec{u}_i, \quad \vec{u}_i \in V \\ L\vec{\varphi}_i = \nu_i \vec{\varphi}_i, \quad \vec{\varphi}_i \in (H_0^1(\Omega))^3 \quad \forall i \geq 1. \end{array} \right. \quad (2.7)$$

Since Ω is of class C^2 , we can apply the regularity results for elliptic equations and it follows that $\vec{u}_i \in V \cap (H^2(\Omega))^3$, $\vec{\varphi}_i \in (H_0^1(\Omega))^3 \cap (H^2(\Omega))^3$ and

$$\begin{cases} \|\vec{v}_m(t)\|_{(H^2(\Omega))^3} \leq c(\Omega) |-\Delta \vec{v}_m(t)|, \\ \|\vec{\omega}_m(t)\|_{(H^2(\Omega))^3} \leq c(\Omega) |L\vec{\omega}_m(t)|. \end{cases} \quad (2.8)$$

We multiply now (2.5)₁ with $\lambda_j g_{jm}(t)$ and we add the equalities. It follows:

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|\vec{v}_m(t)\|_0^2 + (\mu + \chi) |-\Delta \vec{v}_m(t)|^2 &= (\vec{f}(t), -\Delta \vec{v}_m(t)) \\ + \chi (\operatorname{curl} \vec{\omega}_m(t) - \Delta \vec{v}_m(t)) - (b(\vec{v}_m(t), \vec{y}(t)), -\Delta \vec{v}_m(t)) \\ - (b(\vec{u}(t), \vec{v}_m(t)), -\Delta \vec{v}_m(t)). \end{aligned} \quad (2.9)$$

For the right-hand side of (2.9) we use the following inequalities (taking into account the regularity of the functions \vec{y} , \vec{u} , (2.8)₁ and the properties of b):

$$\begin{aligned} -(b(\vec{v}_m(t), \vec{y}(t)), -\Delta \vec{v}_m(t)) &\leq \|\vec{v}_m(t)\|_{(L^4(\Omega))^3} \|\nabla \vec{y}(t)\|_{(L^4(\Omega))^9} |-\Delta \vec{v}_m(t)| \\ &\leq c(\Omega) \|\vec{v}_m(t)\|_{(L^4(\Omega))^3} \|\vec{y}(t)\|_0^{1/4} \|\vec{y}(t)\|_{(H^2(\Omega))^3}^{3/4} |-\Delta \vec{v}_m(t)| \\ &\leq \frac{\mu + \chi}{8} |-\Delta \vec{v}_m(t)|^2 + c \|\vec{y}\|_{L^\infty(0,T;V)}^{1/2} \|\vec{v}_m(t)\|_0^2 \|\vec{y}(t)\|_{(H^2(\Omega))^3}^{3/2}, \\ -(b(\vec{u}(t), \vec{v}_m(t)), -\Delta \vec{v}_m(t)) &\leq \|\vec{u}(t)\|_{(L^4(\Omega))^3} \|\nabla \vec{v}_m(t)\|_{(L^4(\Omega))^9} |-\Delta \vec{v}_m(t)| \\ &\leq \frac{\mu + \chi}{8} |-\Delta \vec{v}_m(t)|^2 + c \|\vec{u}(t)\|_{(L^4(\Omega))^3}^2 \|\vec{v}_m(t)\|_0^{1/2} \|\vec{v}_m(t)\|_{(H^2(\Omega))^3}^{3/2} \\ &\leq \frac{\mu + \chi}{4} |-\Delta \vec{v}_m(t)|^2 + c \|\vec{u}(t)\|_{(L^4(\Omega))^3}^8 \|\vec{v}_m(t)\|_0^2 \end{aligned}$$

With these inequalities, (2.9) becomes

$$\begin{aligned} \frac{d}{dt} \|\vec{v}_m(t)\|_0^2 + (\mu + \chi) |-\Delta \vec{v}_m(t)|^2 &\leq c(|\vec{f}(t)|^2 + \|\vec{\omega}_m(t)\|_0^2) \\ + A(t) \|\vec{v}_m(t)\|_0^2, \end{aligned} \quad (2.10)$$

where $A(t) = c(\|\vec{y}\|_{L^\infty(0,T;V)}^{1/2} \|\vec{y}(t)\|_{(H^2(\Omega))^3}^{3/2} + \|\vec{u}(t)\|_{(L^4(\Omega))^3}^8)$. Using the regularity of the given functions \vec{u} , \vec{y} it follows the integrability of A on $(0, T)$. We can then integrate (2.10) with respect to t and the boundedness of $\{\vec{\omega}_m\}_m$ in $L^2(0, T; (H_0^1(\Omega))^3)$ and of $\{\vec{v}_{0m}\}_m$ in V lead us to the estimates of \vec{v}_m in $L^\infty(0, T; V)$ and in $L^2(0, T; (H^2(\Omega))^3)$. With similar computations we get the corresponding estimates for $\vec{\omega}_m$.

The last step of the proof is to obtain the estimates of $(\vec{v}'_m, \vec{\omega}'_m)$ in $L^2(0, T; (L^2(\Omega))^3)^2$. The inclusion $(H^{2,1}(\Omega_T))^3 \subset L^2(0, T; (L^2(\Omega))^3)$ being compact, the existence of $(\vec{v}, \vec{\omega})$ follows passing to the limit, on a

subsequence, in (2.5). The uniqueness of the pair $(\vec{v}, \vec{\omega})$ is proved as usual.

Corollary 2.3. *Let $(\vec{v}, \vec{\omega})$ be a strong solution of (2.2). Then $\vec{v} \in (H^{2,1}(\Omega_T))^3 \cap C([0, T]; V)$, $\vec{\omega} \in (H^{2,1}(\Omega_T))^3 \cap C([0, T]; (H_0^1(\Omega))^3)$. Moreover, we have:*

$$\begin{cases} \|\vec{v}\|_{(H^{2,1}(\Omega_T))^3} \leq c, \\ \|\vec{\omega}\|_{(H^{2,1}(\Omega_T))^3} \leq c, \end{cases} \quad (2.11)$$

where the constant c depends on Ω , the constants of the problem, the given fields, \vec{f} , \vec{g} , \vec{v}_0 , $\vec{\omega}_0$ and on $\|\vec{v}\|_{L^8(0,T;(L^4(\Omega))^3)}$.

Proof. We take in (2.3) $\vec{y} = \vec{0}$, $\vec{u} = \vec{v}$, $\vec{\rho} = \vec{0}$ and we apply Theorem 2.2. The regularity $\vec{\omega} \in L^8(0, T; (L^4(\Omega))^3)$ follows from the inclusions $C([0, T]; (H_0^1(\Omega))^3) \subset L^q(0, T; (H_0^1(\Omega))^3) \subset L^q(0, T; (L^4(\Omega))^3)$, $\forall q$.

3. Study of the control problem

In the theory of micropolar fluids a special case appears when the microrotation is constrained by:

$$\vec{\omega} = \operatorname{curl} \vec{v}. \quad (3.1)$$

Indeed, if we introduce (3.1) in (2.1)₁, the micropolar fluid becomes a Navier-Stokes one. The aim of this paper is to control the properties of the fluid by acting on the exterior field \vec{g} . The difficulty is that the correspondence $\vec{g} \mapsto (\vec{v}, \vec{\omega})$ is multivalued in the three dimensional case.

For this reason, we formulate the control problem such that to every optimal control we can associate a strong solution. Since we cannot prove the existence of a strong solution, we have to choose a suitable functional.

We define

$$\begin{aligned} J : L^2(0, T; (L^2(\Omega))^3) \times L^2(0, T; V) \times L^2(0, T; (H_0^1(\Omega))^3) &\mapsto \bar{I}\bar{R} \\ J(\vec{g}, \vec{v}, \vec{\omega}) &= \frac{1}{6} \int_0^T (\|\vec{\omega}(t) - \operatorname{curl} \vec{v}(t)\|)^6 dt, \end{aligned} \quad (3.2)$$

with $(\vec{v}, \vec{\omega})$ a solution of (2.2) corresponding to \vec{g} .

Since we cannot expect that we will be able to prove the coercivity of J , we take the exterior field $\vec{g} \in B_r$, with

$$B_r = \{\vec{\varphi} \in L^2(0, T; (L^2(\Omega))^3) / \|\vec{\varphi}\|_{L^2(0,T;(L^2(\Omega))^3)} \leq r\}.$$

We formulate the control problem in the following way:

$$(CP) \quad \left\{ \begin{array}{l} \text{Minimize } J(\vec{g}, \vec{v}, \vec{\omega}) \text{ when } \vec{g} \in B_r \text{ and the pair} \\ (\vec{v}, \vec{\omega}) \text{ verifies (2.2).} \end{array} \right.$$

We establish next the following important result:

Proposition 3.1. *Let $(\vec{g}, \vec{v}, \vec{\omega}) \in B_r \times L^2(0, T; V) \times L^2(0, T; (H_0^1(\Omega))^3)$ with the properties:*

- a) $J(\vec{g}, \vec{v}, \vec{\omega}) < \infty$,
- b) $(\vec{v}, \vec{\omega})$ is a solution of (2.2), corresponding to \vec{g} .

Then $(\vec{v}, \vec{\omega})$ is the unique strong solution of (2.2).

Proof. Since \vec{v} is divergence free it follows that

$$\|\vec{v}\|_0 = |\operatorname{curl} \vec{v}|. \quad (3.3)$$

Using (3.3) we get the following inequality

$$J(\vec{g}, \vec{v}, \vec{\omega}) \geq \frac{1}{6} \int_0^T (\|\vec{v}(t)\|_0^2 - 2\|\vec{v}(t)\|_0 |\vec{\omega}(t)| + |\vec{\omega}(t)|^2)^3 dt.$$

The hypothesis a) of the proposition together with the above inequality implies that $\vec{v} \in L^6(0, T; V)$. On the other hand, from the known inequality $\|u\|_{L^4(\Omega)} \leq \sqrt{2}|u|^{1/4}\|u\|_0^{3/4} \forall u \in H_0^1(\Omega)$ and from the regularity $\vec{v} \in L^\infty(0, T; H)$ given by Theorem 2.1, it follows

$$\|\vec{v}\|_{L^8(0, T; (L^4(\Omega))^3)} \leq \sqrt{2}\|\vec{v}\|_{L^\infty(0, T; H)}^{1/4}\|\vec{v}\|_{L^6(0, T; V)}^{3/4}. \quad (3.4)$$

Remark 3.2. The hypothesis a) from the previous proposition means that for fixed $\vec{f}, \vec{v}_0, \vec{\omega}_0$, we can find a function $\vec{g} \in B_r$ so that the system (2.2) has a strong solution.

Theorem 3.3. *If $J \neq \infty$, then (CP) has at least a solution.*

Proof. We denote

$$m = \inf\{J(\vec{g}, \vec{v}, \vec{\omega}) / \vec{g} \in B_r, (\vec{v}, \vec{\omega}) \text{ solution for (2.2)}\}.$$

Let $\{(\vec{g}_n, \vec{v}_n, \vec{\omega}_n)\}_n$ be a minimizing sequence. From Proposition 3.1. we obtain that $(\vec{v}_n, \vec{\omega}_n)$ is the unique strong solution of (2.2) corresponding to \vec{g}_n . It follows, from Corollary 2.3, that $\{\vec{v}_n\}_n$ and $\{\vec{\omega}_n\}_n$ are bounded in $(H^{2,1}(\Omega_T))^3$ by a constant depending on the fixed data and on $\|\vec{v}_n\|_{L^8(0, T; (L^4(\Omega))^3)}$, which is bounded, from (3.4), by a constant not depending on n . Since the embedding $H^{2,1}(\Omega_T) \subset L^2(\Omega_T)$ is compact, we can pass to the limit in (2.2) corresponding to \vec{g}_n and we obtain that $(\vec{v}, \vec{\omega})$ is the unique strong solution of (2.2) corresponding to \vec{g} , a weak limit point of $\{\vec{g}_n\}_n$ in $(L^2(0, T; (L^2(\Omega))^3))$. It follows that J is weakly lower semicontinuous, and, hence, the proof is achieved.

4. The optimality system

For obtaining the conditions of optimality, we define

$$A = \{\vec{g} \in L^2(0, T; (L^2(\Omega))^3) / (2.2) \text{ corresponding to } \vec{g} \text{ has a strong solution}\} \quad (4.1)$$

and $I : A \mapsto I\mathbb{R}$,

$$I(\vec{g}) = J(\vec{g}, F(\vec{g})) \quad \forall \vec{g} \in A, \quad (4.2)$$

where $F(\vec{g}) = (\vec{v}_g, \vec{\omega}_g)$, $(\vec{v}_g, \vec{\omega}_g)$ is the unique strong solution of (2.2) corresponding to \vec{g} .

Theorem 4.1. *Let \vec{g}^* be an element of A . Then there exists a neighbourhood \mathcal{U} of \vec{g}^* with $\mathcal{U} \subset A$.*

Proof. We shall use the implicit function theorem. For this purpose we introduce the notation:

$$\begin{cases} M = L^2(0, T; (L^2(\Omega))^3), \\ Y = (H^{2,1}(\Omega_T))^3 \cap C([0, T]; V) \times (H^{2,1}(\Omega_T))^3 \cap C([0, T]; (H_0^1(\Omega))^3), \\ Z = L^2(0, T; H) \times L^2(0, T; (L^2(\Omega))^3) \times V \times (H_0^1(\Omega))^3 \end{cases}$$

and we define the operator $\vec{\Phi} : M \times Y \mapsto Z$ with

$$\begin{cases} \Phi_1(\vec{g}, (\vec{v}, \vec{\omega})) = \vec{v}' + b(\vec{v}, \vec{v}) - (\mu + \chi) \Delta \vec{v} - \chi \operatorname{curl} \vec{\omega} - \vec{f}, \\ \Phi_2(\vec{g}, (\vec{v}, \vec{\omega})) = j\vec{\omega}' + jb(\vec{v}, \vec{\omega}) - \gamma \Delta \vec{\omega} - (\alpha + \beta) \nabla(\operatorname{div} \vec{\omega}) \\ \quad + 2\chi \vec{\omega} - \chi \operatorname{curl} \vec{v} - \vec{g} \\ \Phi_3(\vec{g}, (\vec{v}, \vec{\omega})) = \vec{v}(0) - \vec{v}_0, \\ \Phi_4(\vec{g}, (\vec{v}, \vec{\omega})) = \vec{\omega}(0) - \vec{\omega}_0. \end{cases} \quad (4.3)$$

It is obvious that $\vec{\Phi}(\vec{g}^*, (\vec{v}^*, \vec{\omega}^*)) = \vec{0}$. The boundedness and the uniform continuity of the operators $\vec{\Phi}$ and $\frac{\partial \vec{\Phi}(\vec{g}, (\vec{v}, \vec{\omega}))}{\partial(\vec{v}, \vec{\omega})} : Y \mapsto Z$ being easy to obtain, it remains to prove the inversability of the operator $\frac{\partial \vec{\Phi}(\vec{g}^*, (\vec{v}^*, \vec{\omega}^*))}{\partial(\vec{v}, \vec{\omega})}$. The computations give

$$\left\langle \frac{\partial \Phi_i(\vec{g}^*, (\vec{v}^*, \vec{\omega}^*))}{\partial(\vec{v}, \vec{\omega})}, (\vec{v}, \vec{\omega}) \right\rangle = E_i((\vec{v}^*, \vec{\omega}^*), (\vec{v}, \vec{\omega})), \quad i = 1, \dots, 4,$$

where

$$\begin{cases} E_1((\vec{v}^*, \vec{\omega}^*), (\vec{v}, \vec{\omega})) = \vec{v}' + b(\vec{v}, \vec{v}^*) + b(\vec{v}^*, \vec{v}) - (\mu + \chi) \Delta \vec{v} - \chi \operatorname{curl} \vec{\omega}, \\ E_2((\vec{v}^*, \vec{\omega}^*), (\vec{v}, \vec{\omega})) = j\vec{\omega}' + jb(\vec{v}, \vec{\omega}^*) + jb(\vec{v}^*, \vec{\omega}) - \gamma \Delta \vec{\omega} \\ \quad - (\alpha + \beta) \nabla(\operatorname{div} \vec{\omega}) + 2\chi \vec{\omega} - \chi \operatorname{curl} \vec{v}, \\ E_3((\vec{v}^*, \vec{\omega}^*), (\vec{v}, \vec{\omega})) = \vec{v}(0), \\ E_4((\vec{v}^*, \vec{\omega}^*), (\vec{v}, \vec{\omega})) = \vec{\omega}(0). \end{cases} \quad (4.4)$$

It is now obvious that the inversability of $\frac{\partial \vec{\Phi}(\vec{g}^*, (\vec{v}^*, \vec{\omega}^*))}{\partial(\vec{v}, \vec{\omega})}$ is equivalent with the existence and the uniqueness of the solution of the system (2.3) with $\vec{g} = \vec{u} = \vec{v}^*$, $\vec{\rho} = \vec{\omega}^*$. Since the functions \vec{v}^* , $\vec{\omega}^*$ have the regularity required by Theorem 2.2, we can apply this theorem and the

inversability is obtained. Therefore we can use the implicit function theorem and we find a neighbourhood \mathcal{U} of \vec{g}^* and a function $\vec{F} : \mathcal{U} \mapsto Y$ so that $\vec{\Phi}(\vec{g}, \vec{F}(\vec{g})) = \vec{0}$, $\forall \vec{g} \in \mathcal{U}$. If we denote $\vec{F}(\vec{g}) = (\vec{v}_g, \vec{\omega}_g)$, it follows that $(\vec{v}_g, \vec{\omega}_g)$ is the strong solution of (2.2) corresponding to \vec{g} and the proof is achieved.

The problem (CP) can be written in the form:

$$\begin{cases} \text{Find } \vec{g}^* \in A \cap B_r \text{ such that} \\ I(\vec{g}^*) = \min\{I(\vec{g}) / \vec{g} \in A \cap B_r\}. \end{cases} \quad (4.5)$$

Since we have proved that A is an open set, it follows that for every $\vec{g}_1, \vec{g} \in A \cap B_r$ there exists $\delta_0 \in (0, 1)$ so that $\vec{g}_1 + \delta(\vec{g} - \vec{g}_1) \in A \cap B_r$, $\forall \delta \leq \delta_0$. We are now in a position to prove the differentiability of the functional I .

Proposition 4.2. *The functional I is G-differentiable on $A \cap B_r$ and $\forall \vec{g}, \vec{g}_1$*

$$\begin{cases} (I'(\vec{g}_1), \vec{g} - \vec{g}_1)_{L^2(0,T;(L^2(\Omega))^3)} = \\ \int_0^T E_0^2(t) (\operatorname{curl}(\vec{v}^*(t) - \vec{v}_1(t)) - (\vec{\omega}^*(t) - \vec{\omega}_1(t)), \operatorname{curl}\vec{v}_1(t) - \vec{\omega}_1(t)) dt, \end{cases} \quad (4.6)$$

where $E_0(t) = |\operatorname{curl}\vec{v}_1(t) - \vec{\omega}_1(t)|^2$, $(\vec{v}_1, \vec{\omega}_1)$ is the unique strong solution of (2.2) corresponding to \vec{g}_1 and $(\vec{v}^*, \vec{\omega}^*) \in (H^{2,1}(\Omega_T))^3 \cap C([0, T]; V) \times (H^{2,1}(\Omega_T))^3 \cap C([0, T]; (H_0^1(\Omega))^3)$ is the unique solution for the system:

$$\begin{cases} (\vec{v}^{*\prime}(t), \vec{z}) + (b(\vec{v}^*(t), \vec{v}_1(t)), \vec{z}) + (b(\vec{v}_1(t), \vec{v}^*(t) - \vec{v}_1(t)), \vec{z}) \\ + (\mu + \chi)((\vec{v}^*(t), \vec{z}))_0 - \chi(\operatorname{curl} \vec{\omega}^*(t), \vec{z}) = (\vec{f}(t), \vec{z}) \quad \forall \vec{z} \in V, \\ j(\vec{\omega}^{*\prime}(t), \vec{\eta}) + j(b(\vec{v}^*(t), \vec{\omega}_1(t)), \vec{\eta}) + j(b(\vec{v}_1(t), \vec{\omega}^*(t) - \vec{\omega}_1(t)), \vec{\eta}) \\ + \gamma((\vec{\omega}^*(t), \vec{\eta}))_0 + (\alpha + \beta)(\operatorname{div} \vec{\omega}^*(t), \operatorname{div} \vec{\eta}) + 2\chi(\vec{\omega}^*(t), \vec{\eta}) \\ - \chi(\operatorname{curl} \vec{v}^*(t), \vec{\eta}) = (\vec{g}(t), \vec{\eta}) \quad \forall \vec{\eta} \in (H_0^1(\Omega))^3, \\ \vec{v}^*(0) = \vec{v}_0, \vec{\omega}^*(0) = \vec{\omega}_0. \end{cases} \quad (4.7)$$

Proof. The existence, the uniqueness and the regularity of $(\vec{v}^*, \vec{\omega}^*)$ follow from Theorem 2.2. The formula (4.6) is obtained with standard computations, so we shall skip the proof.

Corollary 4.3. *If \vec{g}_1 is a solution for the control problem (4.5), then*

$$\int_0^T E_0^2(t) (\operatorname{curl}(\vec{v}^*(t) - \vec{v}_1(t)) - (\vec{\omega}^*(t) - \vec{\omega}_1(t)), \operatorname{curl}\vec{v}_1(t) - \vec{\omega}_1(t)) dt \geq 0. \quad (4.8)$$

The last result of the paper states the optimality conditions satisfied by an optimal control.

Theorem 4.4. *Let \vec{g}_1 be an optimal control. Then, there exist the unique pairs: $(\vec{v}_1, \vec{\omega}_1)$, the strong solution of (2.2) corresponding to \vec{g}_1 , $(\vec{u}_1, \vec{\rho}_1) \in (H^{2,1}(\Omega_T))^3 \cap C([0, T]; V) \times (H^{2,1}(\Omega_T))^3 \cap C([0, T]; (H_0^1(\Omega))^3)$,*

the unique solution of the adjoint system (4.9), written below which satisfy the following optimality system:

the system (2.2) written for \vec{g}_1 ,

$$\left\{ \begin{array}{l} -(\vec{u}'_1(t), \vec{z}) + (b(\vec{z}, \vec{v}_1(t)), \vec{u}_1(t)) - (b(\vec{v}_1(t), \vec{u}_1(t)), \vec{z}) \\ + j(b(\vec{z}, \vec{\omega}_1(t)), \vec{\rho}_1(t)) + (\mu + \chi)((\vec{u}_1(t), \vec{z}))_0 \\ -\chi(\operatorname{curl} \vec{\rho}_1(t), \vec{z}) = E_0^2(t)(\operatorname{curl} \vec{v}_1(t) - \vec{\omega}_1(t), \operatorname{curl} \vec{z}) \quad \forall \vec{z} \in V, \\ -j(\vec{\rho}'_1(t), \vec{\eta}) - j(b(\vec{v}_1(t), \vec{\rho}_1(t)), \vec{\eta}) + \gamma((\rho_1(t), \vec{\eta}))_0 \\ + (\alpha + \beta)(\operatorname{div} \vec{\rho}_1(t), \operatorname{div} \vec{\eta}) + 2\chi(\vec{\rho}_1(t), \vec{\eta}) - \chi(\operatorname{curl} \vec{u}_1(t), \vec{\eta}) \\ = -E_0^2(t)(\operatorname{curl} \vec{v}_1(t) - \vec{\omega}_1(t), \vec{\eta}) \quad \forall \vec{\eta} \in (H_0^1(\Omega))^3, \\ \vec{u}_1(T) = \vec{0}, \quad \vec{\rho}_1(T) = 0, \end{array} \right. \quad (4.9)$$

$$\int_{\Omega_T} \vec{\rho}_1 \cdot (\vec{g} - \vec{g}_1) dx dt \geq 0 \quad \forall \vec{g} \in A \cap B_r. \quad (4.10)$$

Proof. The variational formulation of the adjoint system (4.9) being of the same type with the variational formulation of (2.3) we can apply Theorem 2.2. and we obtain the existence, the uniqueness and the regularity of $(\vec{u}_1, \vec{\rho}_1)$. The inequality (4.10) is derived from (4.8), taking $(\vec{z}, \vec{\eta}) = (\vec{v}^*(t) - \vec{v}_1(t), \vec{\omega}^*(t) - \vec{\omega}_1(t))$ in (4.9) and $(\vec{z}, \vec{\eta}) = (\vec{u}_1(t), \vec{\rho}_1(t))$ in (4.7)-(2.2) corresponding to \vec{g}_1 . Standard computations complete the proof.

References

- [1] Eringen A. C.: Theory of micropolar fluids, *J. Math. Mech.* **16** (1966), 1–18.
- [2] Temam R.: *Navier-Stokes equations*, North-Holland, Amsterdam, 1977.
- [3] Stavre R.: A distributed control problem for micropolar fluids, *Rev. Roum. Math. Pure Appl.* **45** (2000), 353–358.
- [4] Stavre R.: Optimization and numerical approximation for micropolar fluids, *Preprint IMAR* **6** (2000), submitted.
- [5] Stavre R.: The control of the pressure for a micropolar fluid, *Z. angew. Math. Phys.* **53** (2002), 1–11.
- [6] Casas E.: An optimal control problem governed by the evolution Navier-Stokes equations, *Optimal control of viscous flow*, SIAM, Philadelphia (1998), 79–95.
- [7] Casas E.: The Navier-Stokes equations coupled with the heat equation: analysis and control, *Control and Cybernetics* **23** (1994), 605–620.

AN H^∞ DESIGN METHOD FOR FAULT DETECTION AND IDENTIFICATION PROBLEMS

Adrian M. Stoica

University "Politehnica" of Bucharest

Faculty of Aerospace Engineering

Str. Splaiul Independentei, no. 313, Ro-77206

Bucharest, Romania

amstoica@fx.ro

Michael J. Grimble

University of Strathclyde

Graham Hills Building, 50 George Street,

Glasgow G1 1QE, Scotland

m.grimble@eee.strath.ac.uk

Abstract In the present paper an H^∞ type approach is developed in order to solve fault detection and identification problems for linear dynamic systems. It is shown that the design specifications lead to a multi-objectives optimization problem. Necessary and sufficient solvability conditions for these problems are given together with an illustrative numerical example.

Keywords: Fault detection and identification, H^∞ control, matrix inequalities.

1. Introduction

The fault detection and identification (FDI) problem has been intensively investigated over the last three decades. Roughly speaking, the problem consists in designing a system able to detect, based on the measured outputs of a monitored plant, any failure of the actuators or of the sensors that can occur in the plant functioning. This system is often called in the literature *residual generator*. The problem becomes more difficult when the influence of certain disturbances is taken into account. In this case the residual generator must "distinguish" between the

failure occurrence and the disturbance influence in order to avoid false alarm signals. A more refined statement requires not only to detect a failure but to indicate what failure occurred. This design specification is known as *fault identification* requirement. Some early statements of the problem and corresponding results are given for instance in [1], [8]. Since then, a large number of methods have been proposed to solve FDI problems. Some of them use model-based techniques which include observer-based methods, factorization methods, fuzzy logic and neural control (e.g. [4], [5], [9], [10]). Other methods as are the eigen-structure assignment approaches are based on the properties of some structural invariant subspaces associated with the linear model of the monitored plant (e.g. [2], [3], [11], [12]). The H^2 and H^∞ norms have been also used in the optimization problems arising in the design of the residual generator ([6], [10], [13] and their references).

The present paper describes an H^∞ type method to solve FDI problems. The case when all faults can simultaneously occur at any moment is considered. It is shown that the specific design objectives can be expressed as γ -attenuation conditions. In contrast with the well-known H^∞ control problems, these γ -attenuation conditions cannot be expressed in function of lower linear fractional transformation and therefore the design methodology of the residual generator is different. The paper is organized as follows: in Section 2 the FDI problem is stated and the specific design objectives are transformed in γ -attenuation conditions. Section 3 includes the main result providing necessary and sufficient conditions for the solvability of the FDI problem and the design algorithm of the residual generator. Finally, in Section 4 an illustrative numerical example is presented.

2. Problem statement

Consider the following linear dynamic model of the monitored plant:

$$\begin{aligned}\dot{x} &= Ax + B_1 f + B_2 d \\ y &= Cx + D_1 f + D_2 d\end{aligned}\tag{1}$$

where $x \in \mathbb{R}^n$ is the state variable, $y \in \mathbb{R}^p$ denotes the measured output vector, $f \in \mathbb{R}^{m_f}$ is the fault vector and $d \in \mathbb{R}^{m_d}$ stands for the disturbance signal vector. It is assumed that $f \in L^2([0, \infty), \mathbb{R}^{m_f})$ and $d \in L^2([0, \infty), \mathbb{R}^{m_d})$, where $L^2(\cdot, \cdot)$ denotes the Lebesgue space of square integrable functions. It is also assumed that the system (1) is stable. This is a natural assumption indicating that the monitored plant has been previously stabilized by a corresponding control system which dynamics is included in the state space representation (1). The design problem consists in determining an H^∞ residual controller of n_k order

with the state equations:

$$\begin{aligned}\dot{x}_k &= A_k x_k + B_k y \\ r &= C_k x_k + D_k y\end{aligned}\quad (2)$$

such that the following conditions are accomplished:

- A_k is stable
- Fault detection:

$$\inf_{\omega \in \mathbb{R}} \underline{\sigma}(G_{rf}(j\omega)) > \alpha \quad (3)$$

- Disturbance attenuation:

$$\|G_{rd}(j\omega)\|_\infty < \beta, \quad (4)$$

for some $\alpha > \beta > 0$, where $G_{rf} := KG_{yf}$, $G_{rd} := G_{yd}$ are the transfer functions from f and d to r , respectively, $\underline{\sigma}(\cdot)$ denotes the minimal singular value, and $\|\cdot\|_\infty$ stands for the H^∞ norm.

Another important design objective in the residual generator design is the *fault identification*. This condition requires to take the residual vector r with the same dimension m_f as the fault vector, that is the matrix $D_k D_1$ is square. Moreover, $D_k D_1$ must be invertible; indeed, if $D_k D_1$ is not invertible then for $\omega \rightarrow \infty$ the fault detection condition (3) cannot be fulfilled. The invertibility condition of $D_k D_1$ requires D_1 to be a full rank column matrix.

Remark 1 In applications where D_1 is not left invertible one can consider condition (3) on some specified finite frequency interval $[\underline{\omega}_f, \bar{\omega}_f]$. Introducing an anticausal filter $Q(s)$ such that the output $z(s) = Q(s)y(s)$ satisfies the condition:

$$\sigma(G_{zf}(j\omega)) = \sigma(G_{yf}(j\omega)), \quad \forall \omega \in [\underline{\omega}_f, \bar{\omega}_f]$$

$$\sigma(G_{zf}(j\omega)) \geq \inf_{\underline{\omega}_f \leq \omega \leq \bar{\omega}_f} \underline{\sigma}(G_{yf}(j\omega)), \quad \forall \omega \notin [\underline{\omega}_f, \bar{\omega}_f]$$

one can obtain an equivalent optimization problem in which $G_{zf}(j\omega)$ is invertible on the frequency domain of interest.

Taking into account the above remark, for the further developments the case when D_1 is left invertible will be considered. Then, based on the invertibility assumption of $G_{rf}(j\omega)$, the fault detection condition (3) can be rewritten as:

$$\inf_{\omega \in \mathbb{R}} \underline{\sigma}[(G_{rf}(j\omega))] = \frac{1}{\sup_{\omega \in \mathbb{R}} \overline{\sigma}[G_{rf}^{-1}(j\omega)]} > \alpha,$$

or equivalently,

$$\|G_{rf}^{-1}(j\omega)\|_\infty < \frac{1}{\alpha}. \quad (5)$$

3. Main result

In this section necessary and sufficient conditions for the existence of a stable residual generator satisfying the (3) and (4) are derived. Condition (4) is an H^∞ norm optimization problem but it differs from the usual H^∞ control problems since $G_{rf}^{-1} = (KG_{yf})^{-1}$ cannot be expressed as a lower linear fractional transformation. Therefore condition (5) requires a specific analysis. Thus, since $G_{rf} = KG_{yf}$, based on the realizations (1) and (2) of G_{yf} and K respectively, direct computations give that $G_{rf}^{-1}(s)$ has the realization (A_i, B_i, C_i, D_i) , where:

$$\begin{aligned} A_i &= \begin{bmatrix} A - B_1(D_k D_1)^{-1} D_k C & -B_1(D_k D_1)^{-1} C_k \\ B_k(I - D_1(D_k D_1)^{-1} D_k)C & A_k - B_k D_1(D_k D_1)^{-1} C_k \end{bmatrix}, \\ B_i &= \begin{bmatrix} B_1 \\ B_k D_1 \end{bmatrix} (D_k D_1)^{-1}, \\ C_i &= -(D_k D_1)^{-1} [D_k C \quad C_k] \\ D_i &= (D_k D_1)^{-1}. \end{aligned}$$

Then using the Bounded Real Lemma, in linear matrix inequality (LMI) form, it follows that the equivalent fault detection condition (5) is accomplished if and only if it exists a symmetric matrix $X > 0$ with dimensions $(n + n_k) \times (n + n_k)$ such that:

$$\begin{bmatrix} A_i^T X + X A_i & X B_i & C_i^T \\ B_i^T X & -\frac{1}{\alpha} I & D_i^T \\ C_i & D_i & -\frac{1}{\alpha} I \end{bmatrix} < 0 .$$

Considering the partition of X :

$$X = \begin{bmatrix} R & M \\ M^T & \tilde{R} \end{bmatrix},$$

with $R \in \mathbb{R}^{n \times n}$, for $D_k = U D_1^+$ with U invertible, the above inequality becomes:

$$\begin{bmatrix} \mathcal{L}_{11} & \mathcal{L}_{12} & \mathcal{L}_{13} & \mathcal{L}_{14} \\ \mathcal{L}_{12}^T & \mathcal{L}_{22} & \mathcal{L}_{23} & \mathcal{L}_{24} \\ \mathcal{L}_{13}^T & \mathcal{L}_{23}^T & -\frac{1}{\alpha} I_{m_f} & \mathcal{L}_{34} \\ \mathcal{L}_{14}^T & \mathcal{L}_{24}^T & \mathcal{L}_{34}^T & -\frac{1}{\alpha} I_{m_f} \end{bmatrix} < 0 \quad (6)$$

with

$$\begin{aligned} \mathcal{L}_{11} &= (A - B_1 D_1^+ C)^T R + R (A - B_1 D_1^+ C) + M B_k (I - D_1 D_1^+) C \\ &\quad + C^T (I - D_1 D_1^+)^T B_k^T M^T \end{aligned}$$

$$\begin{aligned} \mathcal{L}_{12} &= (A - B_1 D_1^+ C)^T M + C^T (I - D_1 D_1^+)^T B_k^T \tilde{R} \\ &\quad - (R B_1 + M B_k D_1) (D_k D_1)^{-1} C_k \end{aligned}$$

$$\begin{aligned}
\mathcal{L}_{13} &= (RB_1 + MB_k D_1) (D_k D_1)^{-1} \\
\mathcal{L}_{14} &= -C^T D_1^{+T} \\
\mathcal{L}_{22} &= A_k^T \tilde{R} + \tilde{R} A_k - (MB_1 + \tilde{R} B_k D_1) (D_k D_1)^{-1} C_k \\
&\quad - C_k^T (D_k D_1)^{-T} (MB_1 + \tilde{R} B_k D_1)^T \\
\mathcal{L}_{23} &= (M^T B_1 + \tilde{R} B_k D_1) (D_k D_1)^{-1} \\
\mathcal{L}_{24} &= -C_k^T (D_k D_1)^{-T} \\
\mathcal{L}_{34} &= (D_k D_1)^{-T}
\end{aligned}$$

where $(\cdot)^{-T}$ denotes the transpose of the inverse and $(\cdot)^+$ is the pseudo-inverse of (\cdot) . Further one can set

$$B_k = \begin{cases} \begin{bmatrix} I_p \\ 0 \end{bmatrix} & \text{if } p \geq n_k, \text{ or} \\ [I_{n_k} \ 0] & \text{if } n_k < p. \end{cases} \quad (7)$$

This setting of B_k does not limit the generality of the problem. Indeed, since B_k is not full rank, it exists a small enough perturbation making it full rank and still satisfying the inequality (6). Taking into account the fault identification design objective with simultaneous occurrence of faults, a natural additional condition is

$$D_k D_1 = \delta I_{m_f} \quad (8)$$

with a specified scalar $\delta > 0$. Then direct algebraic computations show that condition (6) is equivalent with:

$$Z_1 + P_1^T \Omega Q_1 + Q_1^T \Omega^T P_1 < 0 \quad (9)$$

with

$$\begin{aligned}
Z_1 &= \begin{bmatrix} Z_{11} & Z_{12} \\ \tilde{A}^T R + R \tilde{A} + \tilde{C}^T M^T + M \tilde{C} & \tilde{A}^T M + \tilde{C}^T \tilde{R} \\ M^T \tilde{A} + \tilde{R} \tilde{C} & 0 \\ B_1^T R + D_1^T B_k^T M^T & B_1^T M + D_1^T B_k^T \tilde{R} \\ -D_1^T C & 0 \end{bmatrix}, \\
Z_{11} &= \begin{bmatrix} RB_1 + MB_k D_1 & -C^T D_1^{+T} \\ M^T B_1 + \tilde{R} B_k D_1 & 0 \\ -\frac{\delta^2}{\alpha} I_{m_f} & -\frac{I_{m_f}}{\alpha} \\ I_{m_f} & -\frac{1}{\alpha} I_{m_f} \end{bmatrix}, \\
Z_{12} &= \begin{bmatrix} M^T & \tilde{R} \\ -\left(B_1^T R + D_1^T B_k^T M^T\right) & -\left(B_1^T M + D_1^T B_k^T \tilde{R}\right) \end{bmatrix}, \\
P_1 &= \begin{bmatrix} 0 & 0 \\ 0 & -I_{m_f} \end{bmatrix}, \\
Q_1 &= \begin{bmatrix} 0 & I_{n_k} & 0 & 0 \end{bmatrix}, \\
\Omega &= \begin{bmatrix} A_k \\ \tilde{C}_k \end{bmatrix}
\end{aligned} \quad (10)$$

where the following notations have been used:

$$\begin{aligned}\tilde{A} &:= A - B_1 D_1^+ C \\ \tilde{C} &:= B_k \left(I - D_1 D_1^+ \right) C \\ \tilde{C}_k &:= (D_k D_1)^{-1} C_k\end{aligned}\quad (11)$$

In the virtue of the Projection Lemma ([11]), there exists Ω satisfying (9) if and only if:

$$W_{P_1}^T Z_1 W_{P_1} < 0 \quad (12)$$

$$W_{Q_1}^T Z_1 W_{Q_1} < 0 \quad (13)$$

where W_{P_1} and W_{Q_1} are bases of the null spaces of P_1 and Q_1 , respectively. Considering the partition of X^{-1} :

$$X^{-1} = \begin{bmatrix} S & N \\ N^T & \tilde{S} \end{bmatrix} > 0$$

with $S \in \mathbb{R}^{n \times n}$, direct algebraic computations show that the above conditions are equivalent with:

$$AS + SA^T - \gamma_1 B_1 B_1^T < 0 \quad (14)$$

and:

$$\begin{bmatrix} \tilde{A}^T R + R \tilde{A} + \tilde{C}^T M^T + M \tilde{C} & RB_1 + MB_k D_1 & -C^T D_1^{+T} \\ B_1^T R + D_1^T B_k^T M^T & -\frac{\delta^2}{\alpha} I_{m_f} & I_{m_f} \\ -D_1^+ C & I_{m_f} & -\frac{1}{\alpha} I_{m_f} \end{bmatrix} < 0, \quad (15)$$

respectively.

Remark 2 Inspecting the 2×2 block matrix in the right lower corner of the matrix in (15) it follows that $\alpha < \delta$.

Recall that $X > 0$ and thus:

$$\begin{bmatrix} R & M \\ M^T & \tilde{R} \end{bmatrix} > 0. \quad (16)$$

Moreover, R and S are related through the condition $XX^{-1} = I$, which leads to the following additional conditions:

$$\begin{bmatrix} R & I_n \\ I_n & S \end{bmatrix} \geq 0 \quad (17)$$

and

$$\text{rank} \left(\begin{bmatrix} R & I_n \\ I_n & S \end{bmatrix} \right) \leq n + n_k. \quad (18)$$

Therefore the fault detection condition (5) is fulfilled if and only if there exist the matrices R, M, \tilde{R} and S satisfying the conditions (14)–(18). In [7] it is proved that if $P_1 P_1^T > 0$ and $Q_1 W_{P_1} W_{P_1}^T Q_1^T > 0$ then the set of all Ω satisfying the basic LMI (9) has the parametrisation:

$$\Omega = \Phi_1 + \Phi_2 L \Phi_3, \quad (19)$$

where the matrix $L \in \mathbb{R}^{(n_k+m_f) \times n_k}$ is any matrix with $\|L\| < 1$ and Φ_1, Φ_2, Φ_3 depend on P_1, Z_1 and Q_1 . The condition $P_1 P_1^T > 0$ directly follows from the expression of P_1 in (10). On the other hand $Q_1 W_{P_1} W_{P_1}^T Q_1^T = \tilde{R}^{-1} M^T M \tilde{R}^{-1}$. Then, for M full rank column matrix Ω has the parameterisation (19). Further, applying again the Bounded Real Lemma in LMI form, one obtains that the disturbance attenuation condition (4) is fulfilled if and only if it exists a symmetric matrix $X_R \in \mathbb{R}^{(n+n_k) \times (n+n_k)} > 0$ such that:

$$\begin{bmatrix} A_R^T X_R + X_R A_R & X_R B_R & C_R^T \\ B_R^T X_R & -\beta I_{m_d} & D_R^T \\ C_R & D_R & -\beta I_{m_d} \end{bmatrix} < 0 \quad (20)$$

where (A_R, B_R, C_R, D_R) is a realization of the system KG_{yd} . Performing the partition

$$X_R = \begin{bmatrix} X_1 & X_2 \\ X_2^T & X_3 \end{bmatrix} > 0, \quad (21)$$

with $X_1 \in \mathbb{R}^{(n \times n)}$, direct algebraic computations show that condition (20) can be rewritten in the equivalent form:

$$Z_2 + P_2^T \Omega Q_2 + Q_2^T \Omega^T P_2 < 0 \quad (22)$$

which, based on the parameterisation (19) of L leads to:

$$\mathcal{N}(X, L) = \begin{bmatrix} \mathcal{N}_{11} & \mathcal{N}_{12} & \mathcal{N}_{13} & \mathcal{N}_{14} \\ \mathcal{N}_{12}^T & \mathcal{N}_{22} & \mathcal{N}_{23} & \mathcal{N}_{24} \\ \mathcal{N}_{13}^T & \mathcal{N}_{23}^T & -\beta I_{m_d} & \mathcal{N}_{34} \\ \mathcal{N}_{14}^T & \mathcal{N}_{24}^T & \mathcal{N}_{34}^T & \mathcal{N}_{44} \end{bmatrix} < 0 \quad (23)$$

with

$$\begin{aligned}
 \mathcal{N}_{11}(X) &= A^T X_1 + X_1 A + C^T B_k^T X_2^T + X_2 B_k C \\
 \mathcal{N}_{12}(X, L) &= A^T X_2 + C^T B_k^T X_3 + X_2 (\Phi_{11} + \Phi_{21} L \Phi_3) \\
 \mathcal{N}_{13}(X) &= X_1 B_2 + X_2 B_k D_2 \\
 \mathcal{N}_{14} &= C^T D_1^{+T} \\
 \mathcal{N}_{22}(X, L) &= X_3 (\Phi_{11} + \Phi_{21} L \Phi_3) + (\Phi_{11} + \Phi_{21} L \Phi_3)^T X_3 \\
 \mathcal{N}_{23}(X) &= X_2^T B_2 + X_3 B_k D_2 \\
 \mathcal{N}_{24}(L) &= \Phi_{12}^T + \Phi_3^T L^T \Phi_{22}^T \\
 \mathcal{N}_{34} &= D_2^T D_1^{+T} \\
 \mathcal{N}_{44} &= -\frac{\beta}{\delta^2} I_{m_f}
 \end{aligned} \tag{24}$$

Φ_{ij} , $i, j = 1, 2$ being given by the partition

$$[\Phi_1 \quad \Phi_2] = \begin{bmatrix} \Phi_{11} & \Phi_{21} \\ \Phi_{12} & \Phi_{22} \end{bmatrix}, \quad \Phi_{11} \in \mathbb{R}^{n_k \times n_k}, \Phi_{21} \in \mathbb{R}^{n_k \times (n_k + m_f)}.$$

Based on the developments above, one proves the following result:

Theorem Given $\alpha > \beta > 0$, the following are equivalent:

- (i) There exists an n_k -order stable residual generator with $D_k = \delta D_1^+$ satisfying the conditions (3) and (4);
- (ii) There exist the symmetric matrices $X_1 \in \mathbb{R}^{(n \times n)}$, $X_3 \in \mathbb{R}^{(n_k \times n_k)}$ and the rectangular matrices $X_2 \in \mathbb{R}^{(n \times n_k)}$, $L \in \mathbb{R}^{(n_k + m_f) \times n_k}$ satisfying (21), (23) and

$$\begin{bmatrix} I_{n_k + m_f} & L \\ L^T & I_{n_k} \end{bmatrix} > 0. \tag{25}$$

If these conditions are satisfied then the matrices A_k and C_k are given by:

$$\begin{bmatrix} A_k \\ C_k \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & \delta I_{m_f} \end{bmatrix} (\Phi_1 + \Phi_2 L \Phi_3). \tag{26}$$

The algorithm to determine an n_k order residual generator solving the fault detection problem (3), (4) is the following:

Step 1 Solve the system of matrix inequalities (14)–(18) with respect to R, S, \tilde{R} and Δ ;

Step 2 Determine the parameterisation (19) of the set of all Ω with the free parameter L , $\|L\| < 1$;

Step 3 Solve the system of matrix inequalities (21), (23), (25) with respect to X_1, X_2, X_3 and L ;

Step 4 Introduce the computed L into (26) to obtain A_k and C_k .

4. A numerical example

Consider the linearized longitudinal dynamics of an aircraft subjected to wind disturbances:

$$\begin{aligned}\dot{x} &= Ax + B_\omega \omega + B_\delta \delta \\ y &= Cx,\end{aligned}$$

where $x \in \mathbb{R}^5$ the state vector including the longitudinal body axis velocity u , the normal body axis velocity w , the pitch rate q , the pitch angle θ and the wind gust w_g . The control variable δ is the elevon deflection angle and ω is the white noise input with unit spectral density generating the wind gust disturbance w_g . The measured outputs vector y includes: the pitch rate q , the angle of attack α , the normal acceleration a_z and the longitudinal acceleration a_x . For the F16XL fighter at the altitude $h = 10000\text{ft}$ and the speed Mach=0.9 the matrices of the state-space realization are ([6]):

$$\begin{aligned}A &= \begin{bmatrix} -0.0674 & 0.0430 & -0.8886 & -0.5587 & 0.0430 \\ 0.0205 & -1.4666 & 16.5800 & -0.0299 & -1.4666 \\ 0.1377 & -1.6788 & -0.6819 & 0 & -1.6788 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1.1948 \end{bmatrix}, \\ B_\omega &= \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1.57 \end{bmatrix}, \quad B_\delta = \begin{bmatrix} -0.1672 \\ -1.5172 \\ -9.7842 \\ 0 \\ 0 \end{bmatrix}, \\ C &= \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0.0591 & 0 & 0 & 0.0591 \\ 0.0139 & 1.0517 & 0.1495 & -0.0299 & 0 \\ -0.0677 & 0.0431 & 0.0171 & 0 & 0 \end{bmatrix}.\end{aligned}$$

The two faults are considered in this example, namely the elevon and the normal acceleration accelerometer failures, respectively. Then $m_f = 2$ and, corresponding to representation (1),

$$B_1 = [B_\delta \ 0_{5 \times 1}], \quad B_2 = B_\omega, \quad D_1 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad D_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Since D_1 is not full rank the fault detection condition (3) cannot be accomplished for high frequency failure signals. Then G_{yf} have been scaled

according with Remark 1. Applying the theoretical results presented in Section 3 an H^∞ residual generator of order $n_k = 5$ was obtained. The nonlinear matrix inequality (23) has been solved by an iterative *centering algorithm*. The time responses corresponding to individual and simultaneous failures have been determined. The numerical simulations are performed for unit fault signals occurring at $t = 3$ seconds and they are presented in Figures 1, 2 and 3, respectively.

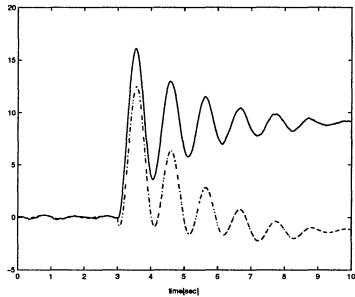


Figure 1. The step responses of the residuals ($r_1 -$, $r_2 - \cdot$): the first fault occurs

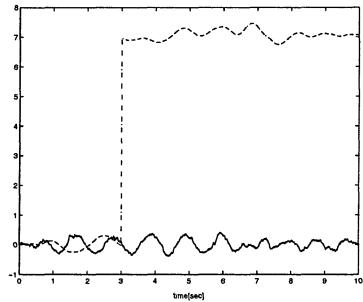


Figure 2. The step responses of the residuals ($r_1 -$, $r_2 - \cdot$): the second fault occurs

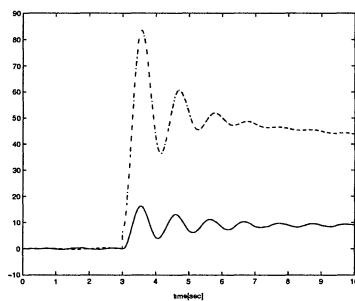


Figure 3. The step responses of the residuals ($r_1 -$, $r_2 - \cdot$): the case when the two faults simultaneously occur

5. Conclusions

An H^∞ type method to solve fault detection and identification problems has been considered. Necessary and sufficient solvability conditions have been derived in terms of some systems of matrix inequalities. The proposed approach allows to develop similar design methods in the discrete-time, stochastic and hybrid representations of the monitored plants.

References

- [1] Beard, R.U. "Failure accommodation in linear-systems through self reorganization", Ph.D. Thesis, Department of Aeronautics and Astronautics, MIT, Cambridge, MA, 1971.
- [2] Chen, R.H., Speyer J.L. "Optimal Stochastic Multiple-Fault Detection Filter" Proceedings of the 38th CDC, Phoenix, Arizona, USA, 1999.
- [3] Choi J.W. "A simultaneous assignment methodology of right/left eigenstructures", *IEEE Transactions on Aerospace and Electronic Systems*, 1998, **34**(2), 625-634;
- [4] Frank P.M., Ding X. "Frequency domain approach to optimally robust generation and evaluation for model-based fault diagnosis", *Automatica*, 1994, **30**, 789-904.
- [5] Frank P.M., Ding X. "Survey of robust residual generation and evaluation methods in observer-based fault detection systems", *Journal of Process Control*, 1997, **7**, 403-424.
- [6] Douglas R.K., Speyer J.L. " H_∞ Bounded Fault Detection Filter", *Journal of Guidance, Control and Dynamics*, **22**(1), 129-139, 1999.
- [7] Iwasaki, T., Skelton, R.E., " All controllers for the general H_∞ control problem: LMI existence conditions and state space formulas", *Automatica*, 1994, **30**, pp. 1307-1317.
- [8] Jones H.L. "Failure detection in linear systems", Ph.D. thesis, Department of Aeronautics and Astronautics, MIT, Cambridge, MA, 1973.
- [9] Patton R.J., Chen J. "Parity space approach to model-based fault diagnosis. A tutorial survey and some new results", *Proceedings of IFAC/IMACS Symposium of SAFAPROCESS'91*, Baden-Baden, 1991.
- [10] Patton R.J., Hou M. " H_∞ estimation and robust fault detection", *Proceedings of European Control Conference, Karlsruhe, Germany*, 1999.
- [11] Patton R.J., Chen J. "On eigenstructure assignment for robust fault diagnosis", *International Journal of Robust and Nonlinear Control*, 2000, **10**:1193-1208.
- [12] Saberi A., Stoorvogel A., Sannuti P., Niemann H. "Fundamental problems in fault detection and identification", *International Journal of Robust and Non-linear Control*, 2000, **10**, 1209-1236.
- [13] Sauter D., Rambaux F., Hamelin F. "Robust fault diagnosis in an H_∞ setting", *Proceedings of SAFAPROCESS'97*, 1997, 879-884.
- [14] Skelton R.E., Iwasaki T., Grigoriadis K. *A Unified Algebraic Approach to Linear Control Design*, Taylor & Francis, 1998.

RICCATI EQUATION OF STOCHASTIC CONTROL AND STOCHASTIC UNIFORM OBSERVABILITY IN INFINITE DIMENSIONS

Viorica Mariela Ungureanu

*Faculty of Engineering, "Constantin Brâncusi" University,
Bul Republicii, nr.1, Târgu -Jiu, jud. Gorj, 1400, România
vio@utgjiu.ro*

Abstract We establish that under stabilizability and observability conditions the Riccati equation arising in the stochastic quadratic control problem has a unique uniformly positive bounded solution.

Keywords: Riccati equation, stochastic uniform observability, stabilizability, detectability, uniform controllability

Introduction

G.Da Prato and I. Ichikawa proved in [3] that under stabilizability and detectability conditions the Riccati equation (3.1) arising in the stochastic quadratic control problem has a unique nonnegative bounded solution. We replace the detectability condition with the uniform observability property and we obtain the same conclusion and, moreover, we show that this solution is uniformly positive. So, we generalize the result obtained by T.Morozan in [7] for finite dimensional case. We also prove that our result is distinct from the one of G.Da Prato and I. Ichikawa.

1. Stabilizability, detectability, uniform observability and uniform controllability for linear differential stochastic equations

Let H, U, V be separable real Hilbert spaces. $L(H, V)$ is the Banach space of all bounded linear operators from H into V (if $H = V$ we put $L(H, V) = L(H)$). We denote by \mathcal{H} the subspace of $L(H)$ formed by

all self-adjoint operators. We write $\langle \cdot, \cdot \rangle$ for the inner product and $\|\cdot\|$ for norms of elements and operators. The operator $S \in L(H)$ is called nonnegative ($S \geq 0$) if S is self-adjoint and $\langle Sx, x \rangle \geq 0$ for all $x \in H$. We set $L^+(H) = \{S \in L(H), S \geq 0\}$. For each interval $J \subset \mathbf{R}_+$, we denote by $C_s(J, L(H))$ the space of all strongly continuous mappings $G(t) : J \subset \mathbf{R}_+ \rightarrow L(H)$ and by $C_b(J, L(H))$ the subspace of $C_s(J, L(H))$, which consist of all mappings $G(t)$ such that $\sup_{t \in J} \|G(t)\| < \infty$. If E is a Banach

space we also denote by $C(J, E)$ the space of all continuous mappings $G(t) : J \subset \mathbf{R}_+ \rightarrow E$. We need the following assumption:

- P_1 : a) $A(t)$, $t \in [0, \infty)$ is a closed linear operator on H with constant domain D dense in H .
- b) there exist $M > 0$, $\eta \in (\frac{1}{2}\pi, \pi)$ and $\delta \in (-\infty, 0)$ such that $S_{\delta, \eta} = \{\lambda \in C; |\arg(\lambda - \delta)| < \eta\} \subset \rho(A(t))$, for all $t \geq 0$ and $\|R(\lambda, A(t))\| \leq \frac{M}{|\lambda - \delta|}$ for all $\lambda \in S_{\delta, \eta}$.
- c) there exist numbers $\alpha \in (0, 1)$ and $\tilde{N} > 0$ such that $\|A(t)A^{-1}(s) - I\| \leq \tilde{N}|t - s|^\alpha$, $t \geq s \geq 0$, where we denote by $\rho(A)$, $R(\lambda, A)$ the resolvent set of A and respectively the resolvent of A .

It is known (see [4]) that if P_1 holds, then the family $A(t)$, $t \geq 0$ generates an evolution operator $U(t, s)$. For any $n \in N$ we have $n \in \rho(A(t))$. The operators $A_n(t) = n^2 R(n, A(t)) - nI$ are called the Yosida approximations of $A(t)$. If we denote by $U_n(t, s)$ the evolution operator relative to $A_n(t)$ for each $x \in H$ one has $\lim_{n \rightarrow \infty} U_n(t, s)x = U(t, s)x$ uniformly on any bounded subset of $\{(t, s); t \geq s \geq 0\}$.

Let $(\Omega, F, F_t, t \in [0, \infty), P)$ be a stochastic basis. We consider the stochastic equation

$$dy(t) = A(t)y(t)dt + \sum_{i=1}^m G_i(t)y(t)dw_i(t), \quad y(s) = x \in H,$$

denoted by $\{A; G_i\}$. The family $A(t)$ satisfies the hypothesis P_1 , $G_i \in C_s([0, \infty), L(H))$, $i = 1, \dots, m$ and w_i 's are independent real Wiener processes relative to F_t .

It is known (see [3]) that $\{A; G_i\}$ has a unique mild solution in the space $C([s, T], L^2(\Omega; H))$ that is adapted to F_t ; namely the solution of

$$y(t) = U(t, s)x + \sum_{i=1}^m \int_s^t U(t, r)G_i(r)y(r)dw_i(r). \quad (1.1)$$

Let $y(t, s; x)$ be the mild solution of $\{A; G_i\}$. We have the following definition (see [5] for the autonomous case):

Definition 1 The equation $\{A; G_i\}$ is uniformly exponentially stable if there exist constants $M \geq 1, \omega > 0$ such that $E\|y(t, s; x)\|^2 \leq M e^{-\omega(t-s)} \|x\|^2$ for all $t \geq s \geq 0$ and $x \in H$.

If $C \in C_s([0, \infty), L(H, V))$, we consider the system $\{A, G_i; C\}$ formed by equation $\{A; G_i\}$ and the observation relation $z(t) = C(t)y(t, s, x)$.

Definition 2 [7] The system $\{A, C; G_i\}$ is uniformly observable if there exist $\tau > 0$ and $\gamma > 0$ such that $E \int_s^{s+\tau} \|C(t)y(t, s; x)\|^2 dt \geq \gamma \|x\|^2$ for all $s \in \mathbf{R}_+$ and $x \in H$.

Definition 3 [3] Let $D \in C_b([0, \infty), L(H))$. The system $\{A, D; G_i\}$ is detectable if there exists $L \in C_b([0, \infty), L(H))$ such that $\{A + LD; G_i\}$ is uniformly exponentially stable.

If $B \in C_b([0, \infty), L(U, H))$ and $u \in L^2(\mathbf{R}_+, U)$, we associate to $\{A; G_i\}$ the following stochastic linear control equation, denoted by $\{A, B; G_i\}$:

$$dy(t) = A(t)y(t)dt + B(t)u(t)dt + \sum_{i=1}^m G_i(t)y(t)dw_i(t), \quad y(s) = x.$$

Definition 4 [3] The equation $\{A, B; G_i\}$ is stabilizable if there exists $F \in C_b([0, \infty), L(H, U))$ such that $\{A + BF; G_i\}$ is uniformly exponentially stable.

In the deterministic case it is known (see [6] for the autonomous case) that uniform observability implies detectability. We will prove in the following section that this assertion is not true in the stochastic case.

2. A counter example for the implication "uniform controllability implies stabilizability"

Assume that $H = \mathbf{R}^n, V = \mathbf{R}^p$ and $U = \mathbf{R}^d$ (where \mathbf{R}^n is the real n -dimensional space) and functions A, B, C and G_i are defined on the whole real axis \mathbf{R} and are assumed to be continuous and bounded. By $Y(t, s), t \geq s$ we denote the fundamental random matrix solution associated with the equation $\{A; G_i\}$ (see [8]).

Definition 5 The equation $\{A, B; G_i\}$ is uniformly controllable if there exist $\tau > 0$ and $\gamma > 0$ such that $E \int_{s-\tau}^s Y(s, t)B(t)B^*(t)Y^*(s, t)dt \geq \gamma I$ for all $s \in \mathbf{R}$, where I is the identity operator on H .

Remark 1 If $G_i = 0$ and $Y(s, t)$ is replaced with the evolution operator $U(s, t)$ associated to family $A(t)$, $t \in \mathbf{R}$ we obtain the definition for the uniform controllability of the deterministic equation $\{A, B\}$.

Similarly, the Definition 2 can be extended for all $s \in \mathbf{R}$. So, by corollaries C.1 and C.2 of [8] we deduce that uniform controllability implies (or does not imply) stabilizability if and only if uniform observability implies (or does not imply) detectability. Suppose all the operators are constants, $m=1$ and $G_1=G$. In the time invariant case uniform controllability is called controllability. From Proposition 4.1 in [3], it follows:

Remark 2 If $\{A, B; G\}$ is stabilizable, then there exists a solution in $L^+(H)$ of the Riccati equation

$$A^*K + KA + G^*KG + I - KBB^*K = 0. \quad (2.1)$$

Lemma 1¹ (see [9]) *If the deterministic equation $\{A, B\}$ is controllable (see the Remark 1), then the stochastic equation $\{A, B; G\}$ is controllable.*

Proof. Let us consider the operators $\widetilde{M}, \widetilde{G}: \mathcal{H} \rightarrow \mathcal{H}$, $\widetilde{M}(P)=AP+PA^*$, $\widetilde{G}(K) = GKG^*$. From proposition P.1 in [8], we see that $\{A, B; G\}$ (resp. $\{A, B\}$) is controllable, if and only if there exist $\tau, \gamma > 0$ such that $\int_{s-\tau}^s e^{(s-p)(\widetilde{M}+\widetilde{G})} BB^*dp \geq \gamma I$ (resp. $\int_{s-\tau}^s e^{(s-p)(\widetilde{M})} BB^*dp \geq \gamma I$) for all $s \in \mathbf{R}$. If we consider the Cauchy problem

$$\frac{dR(t)}{dt} = (\widetilde{M} + \widetilde{G})(R(t)), R(s) = BB^*, t \geq s \geq 0$$

we have (by variation constants formula)

$$R(t) = e^{(t-s)\widetilde{M}}(BB^*) + \int_s^t e^{(t-r)\widetilde{M}}\widetilde{G}(R(r))dr.$$

On the other hand $R(t) = e^{(t-s)(\widetilde{M}+\widetilde{G})}(BB^*) \geq 0$.

Since $\widetilde{G}(K) \geq 0$ and $e^{(t-s)\widetilde{M}}(K) \geq 0$ for all $K \in L^+(H)$, we deduce $e^{(t-s)(\widetilde{M}+\widetilde{G})}(BB^*) \geq e^{(t-s)\widetilde{M}}(BB^*)$ for all $t \geq s \geq 0$. From the hypothesis and the last inequality the conclusion follows. ■

Counter-example. (see[9]) Let us consider the stochastic equation $\{A, B; G\}$, where $A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$, $B = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $G = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$. Since $\text{rank}(B, AB) = 2$, $\{A, B\}$ is controllable and, from the previous lemma, we deduce that $\{A, B; G\}$ is controllable. We will prove that $\{A, B; G\}$

¹This result was proved in [9] in the hypothesis that the operators A and G commute, but prof.T.Morozan noticed that this assumption can be removed.

is not a stabilizable equation. Assume by contradiction that $\{A, B; G\}$ is stabilizable. Then, from Remark 2, follows that there exists a solution $K = \begin{pmatrix} x_1 & x_2 \\ x_2 & x_3 \end{pmatrix}$ in $L^+(H)$ of the Riccati equation (2.1), which satisfies the conditions

$$x_1 x_3 \geq x_2^2, x_1 \geq 0. \quad (2.2)$$

For this solution, equation (2.1) is equivalent with the following system:

$$\begin{aligned} (x_1 + x_2)^2 &= 3x_1 + 1, (x_1 + x_2)(x_2 + x_3) = 3x_2, \\ (x_2 + x_3)^2 &= 4x_3 + 1. \end{aligned} \quad (2.3)$$

In order to solve (2.3) the following situations arise : a) if $x_1 + x_2 = 0$ (respectively $x_2 + x_3 = 0$), then we have $x_1 = -1/3$ (respectively $x_3 = -1/4$) and the condition (2.2) does not hold;

b) if $x_1 + x_2 \neq 0$ and $x_2 + x_3 \neq 0$, we obtain $\frac{x_1+x_2}{x_2+x_3} = \frac{3x_1+1}{3x_2} = \frac{3x_2}{4x_3+1}$. Hence it follows successively $9x_2^2 = 12x_1x_3 + 3x_1 + 4x_3 + 1 > 12x_1x_3$ and $9/12x_2^2 > x_1x_3$. From (2.2) we deduce $9/12x_2^2 > x_2^2$ and we get a contradiction. Thus, we proved that there exists a stochastic system, which is controllable and is not stabilizable. Consequently, an uniform observable system it not necessarily a detectable one.

3. Bounded solutions of Riccati equation of stochastic control

In this section assume that P_1 holds, $B \in C_b([0, \infty), L(U, H))$, $B^* \in C_b([0, \infty), L(H, U))$, $C \in C_b([0, \infty), L(H, V))$, $C^*C, G_i \in C_b([0, \infty), L(H))$, $K(t) \in C_b([0, \infty), L^+(U))$ and there exists $\delta_0 > 0$ such that $K(t) \geq \delta_0 I$ for all $t \in [0, \infty)$. Consider the Riccati equation

$$\begin{aligned} P'(s) + A^*(s)P(s) + P(s)A(s) + \sum_{i=1}^m G_i^*(s)P(s)G_i(s) + \\ + C^*(r)C(r) - P(s)B(s)(K(s))^{-1}B^*(s)P(s) = 0 \end{aligned} \quad (3.1)$$

We say that P is a mild solution on an interval J of (3.1)(see [3]), if $P \in C_s(J, L^+(H))$ and if it satisfies

$$\begin{aligned} P(s)x = U^*(t, s)P(t)U(t, s)x + \sum_{i=1}^m \int_s^t U^*(r, s)[G_i^*(r)P(r)G_i(r) + \\ + C^*(r)C(r) - P(r)B(r)(K(r))^{-1}B^*(r)P(r)]U(r, s)xdr \end{aligned} \quad (3.2)$$

for all $s \leq t, s, t \in J$. Moreover, if P is a mild solution on \mathbf{R}_+ of (3.1) and $\sup_{s \in \mathbf{R}_+} \|P(s)\| < \infty$, then P is said to be a *bounded solution*. If

$$\begin{aligned} P'_n(s) + A_n^*(s)P_n(s) + P_n(s)A_n(s) + \sum_{i=1}^m G_i^*(s)P_n(s)G_i(s) + \\ + C^*(r)C(r) - P_n(s)B(s)(K(s))^{-1}B^*(s)P_n(s) = 0 \end{aligned} \quad (3.3)$$

is the approximating equation of (3.1), then we have the following lemma:

Lemma 2 [3] *Assume P_1 holds. Let $0 < T < \infty$ and let $R \in L^+(H)$. Then there exists a unique mild (resp. classical) solution P (resp. P_n) of (3.1) (resp. (3.3)) on $[0, T]$ such that $P(T) = R$ (resp. $P_n(T) = R$) and for each $x \in H$, $P_n(s)x \rightarrow P(s)x$ uniformly on $[0, T]$.*

Assume that (3.1) has a bounded solution $P(s)$ and consider $S(s) = -(K(s))^{-1}B^*(s)P(s)$, $s \geq 0$. We denote $L^s(H) = L^2(\Omega, F_s, P, H)$ and if $\xi \in L^s(H)$ we consider the equation

$$dz(t) = [A(t) + B(t)S(t)]z(t)dt + \sum_{i=1}^m G_i(t)z(t)dw_i(t), z(s) = \xi \quad (3.4)$$

We associate to (3.4) the integral equation:

$$\begin{aligned} z(t) = U(t, s)\xi + \int_s^t U(t, r)B(r)S(r)z(r)dr + \\ + \sum_{i=1}^m \int_s^t U(t, r)G_i(r)z(r)dw_i(r) \end{aligned} \quad (3.5)$$

If $A_n(t)$, $n \in N$ are the Yosida approximations of $A(t)$, then we have the following approximating system of (3.4)

$$\begin{aligned} dz_n(t) = [A_n(t) + B(t)S(t)]z_n(t)dt + \sum_{i=1}^m G_i(t)z_n(t)dw_i(t), \\ z_n(s) = \xi \end{aligned} \quad (3.6)$$

Lemma 3 *Let $0 \leq s \leq T$. If $P(t)$, $t \in \mathbf{R}_+$ is a bounded solution of (3.1), then (3.5) has a unique solution denoted by $z(t, s; x)$, which belongs to $C([s, T], L^2(\Omega; H))$ and is adapted to F_t and the system (3.6) has a unique classical solution z_n . Moreover, $z_n \rightarrow z$ in mean square uniformly on $[s, T]$.*

The unique solution of (3.5) which belongs to $C([s, T], L^2(\Omega; H))$ is called the mild solution of (3.4).

Proof. From Theorem 2.3.1 in [1] it follows that (3.4) and respectively (3.6) have unique mild solutions in $C([s, T], L^2(\Omega; H))$. It is a simple exercise to verify that the mild solution of (3.6) coincides with the strong (classical) solution. We only have to prove that $z_n \rightarrow z$ in mean square uniformly on $[s, T]$, $s \geq 0$. Since $U_n(t, r)x - U(t, r)x \xrightarrow{n \rightarrow \infty} 0$ for every $x \in H$ uniformly on $\{(t, r), T \geq t \geq r \geq 0\}$, we deduce, from the uniform boundedness theorem, that there exists $\tilde{M}_T > 0$ such as $\|U_n(t, r) - U(t, r)\| \leq \tilde{M}_T$ for all $0 \leq r \leq t \leq T$ and $n \in N$. If $M_T = \sup_{0 \leq r \leq t \leq T} \|U(t, r)\|$, $\tilde{G}_i = \sup_{r \in [0, T]} \|G_i(r)\|$, $i = 1, \dots, m$ and

$\tilde{D} = \sup_{r \in [0, T]} \|B(r)S(r)\|$, then it is not very difficult to see that we have

$$\begin{aligned} E \|z_n(t) - z(t)\|^2 &\leq (2m+3)\{\|U_n(t,s)x - U(t,s)x\|^2 + \\ &+ \widehat{N} \int_s^t E \|z_n(r) - z(r)\|^2 dr + \int_s^t E \|[U_n(t,r) - U(t,r)]D(r)z(r)\|^2 dr + \quad (3.7) \\ &+ \sum_{i=1}^m \int_s^t E \|[U_n(t,r) - U(t,r)]G_i(r)z(r)\|^2 dr\}, \end{aligned}$$

where \widehat{N} is a constant depending on $\widetilde{M}_T, M_T, \widetilde{D}$ and $\widetilde{G}_i, i = 1, \dots, m$. Denote the last two integrals of the previous inequality by $I_{1,n}$ and $I_{2,n}$, respectively. It is easy to see that if we prove that $I_{1,n} \xrightarrow{n \rightarrow \infty} 0$ (resp. $I_{2,n} \xrightarrow{n \rightarrow \infty} 0$) uniformly with respect to t , then for every $\varepsilon > 0$ there exists $n_\varepsilon \in N$ such as $\|U_n(t,s)x - U(t,s)x\|^2 + I_{1,n} + I_{2,n} < \varepsilon$ for all $n \in N$, $n \geq n_\varepsilon$. Using this relation in (3.7) and Gronwall's lemma it follows the conclusion. We prove only the statement for $I_{1,n}$, because the proof for $I_{2,n}$ goes on similarly. Since $\phi : [s, T] \rightarrow L^2(\Omega, H)$ $\phi(r) = D(r)z(r)$ is continuous, then $\phi([s, T]) = \Delta \subset L^2(\Omega, H)$ is compact. Let fix $r \in [s, T]$ and $\alpha_n(r) = \sup_{(t,\xi) \in [r,T] \times \Delta} E \|[U_n(t,r) - U(t,r)]\xi\|^2$. It is easy to see that

the map $(t, \xi) \xrightarrow{\Psi} E \|[U_n(t,r) - U(t,r)]\xi\|^2$ is continuous on $[r, T] \times \Delta$ for every $n \in N$. Thus there exists $(t_{n,r}, \xi_{n,r}) \in [r, T] \times \Delta$ such that $\alpha_n(r) = \Psi(t_{n,r}, \xi_{n,r})$. Since r is fixed, we write, by convenience, (t_n, ξ_n) instead of $(t_{n,r}, \xi_{n,r})$ and so, $\alpha_n(r) = E \|[U_n(t_n, r) - U(t_n, r)]\xi_n\|^2$. We will prove that $\alpha_n(r) \xrightarrow{n \rightarrow \infty} 0$. Assume by contradiction that $\alpha_n(r) \xrightarrow{n \rightarrow \infty} 0$. Let $\delta > 0$. Then there exists a subsequence of $\alpha_n(r)$ such as $\alpha_n(r) \geq \delta$. Since $\{t_n\}_{n \in N}, \{\xi_n\}_{n \in N}$ belongs to compact sets, they have convergent subsequences. Consider a subsequence of $\{\alpha_n(r)\}_{n \in N}$ (still denoted $\alpha_n(r)$) such as the corresponding $\{t_n\}$ and $\{\xi_n\}$ converge. We have $\lim_{n \rightarrow \infty} E \|[U_n(t_n, r) - U(t_n, r)]\xi_n\|^2 \leq 5 \lim_{n \rightarrow \infty} \{E \|[U_n(t_n, r)(\xi_n - \xi)]\|^2 +$

$$\begin{aligned} &+ E \|[U_n(t_n, r) - U(t_n, r)]\xi\|^2 + 2E \|[U(t_n, r) - U(t, r)]\xi\|^2 + \\ &+ E \|[U(t_n, r)(\xi_n - \xi)]\|^2\} = 0. \end{aligned}$$

We obtain, $\lim_{n \rightarrow \infty} \alpha_n(r) = 0 \geq \delta$ and we deny the hypothesis; hence $\lim_{n \rightarrow \infty} \alpha_n(r) = 0$. Consider the map $\Phi_n(t, r, \xi) : [s, T] \times [s, T] \times \Delta \rightarrow \mathbf{R}_+$,

$$\Phi_n(t, r, \xi) = E \|[U_n(t, r) - U(t, r)]\xi\|^2 \chi_{\{(t,r), s \leq r \leq t \leq T\}}(t, r).$$

Since $(t, \xi) \rightarrow \Phi_n(t, r, \xi)$ (resp. $r \rightarrow \Phi_n(t, r, \xi)$) is continuous on $[s, T] \times \Delta$ (resp. on $[s, T]$), $\sup_{(t,\xi) \in [s,T] \times \Delta} \Phi_n(t, r, \xi) \leq \alpha_n(r)$ and $[s, T] \times \Delta$ is separable, it follows that $r \rightarrow \sup_{(t,\xi) \in [s,T] \times \Delta} \Phi_n(t, r, \xi)$ is Borel measurable. Then

$$I_{1,n} = \int_s^T \Phi_n(t, r, \xi) dr \leq \int_s^T \sup_{(t,\xi) \in [s,T] \times \Delta} \Phi_n(t, r, \xi) dr \xrightarrow{n \rightarrow \infty} 0 \text{ by bounded}$$

convergence theorem ($\Phi_n(t, r, \xi) \leq \widetilde{M}_T \sup_{\xi \in \Delta} E \|\xi\|^2 < \infty$). ■

We introduce the following hypothesis:

P_2 : The evolution operator $U(t, s)$ has an exponentially growth that is, there exist M_0 and ω positive constants such that $\|U(t, s)\| \leq M_0 e^{\omega(t-s)}$.

Taking mean square in (3.5) and using P_2 and Gronwall's inequality we obtain:

Lemma 4 Assume P_2 holds. If $z(t, s; x)$ is the mild solution of (3.4), then there exists a continuous function $\phi : \mathbf{R}_+ \rightarrow \mathbf{R}_+$ such that

$$E \|z(t, s; x)\|^2 \leq \phi(t-s) \|x\|^2 \text{ for all } x \in H \text{ and } 0 \leq s \leq t.$$

Definition 6 A self-adjoint solution of (3.1) is called stabilizing for $\{A; G_i\}$ if $\{A + BS; G_i\}$ is uniformly exponentially stable, where $S(t) = -(K(t))^{-1} B^*(t) P(t)$.

Theorem 1 Assume that $\{A, G_i; C\}$ is uniformly observable and P_2 holds. If $P(t)$ is a nonnegative bounded solution of (3.1) then

a) there exists $\delta > 0$ such that $P(t) \geq \delta I$ for all $t \in [0, \infty)$ (P is uniformly positive on \mathbf{R}_+);

b) P is a stabilizing solution (for $\{A; G_i\}$).

Proof. The main idea is the one in [7]. Let $P(t)$ be a nonnegative bounded solution of the Riccati equation (3.1) and γ and τ as in Definition 2. If $z(t, s; x)$ is the mild solution of (3.4) then we introduce the linear nonnegative operator $Q(s)$, which satisfies the relation

$$\begin{aligned} \langle Q(s)x, x \rangle &= E \int_s^{s+\tau} \|C(t)z(t, s; x)\|^2 + \\ &\quad + \langle S^*(t)K(t)S(t)z(t, s; x), z(t, s; x) \rangle dt. \end{aligned} \tag{3.8}$$

The above relation defines a unique linear nonnegative operator $Q(s)$.

We will prove that $\mathcal{I} = \inf\{\langle Q(s)x, x \rangle, s \geq 0, x \in H, \|x\| = 1\} > 0$.

Assume by contradiction that $\mathcal{I} = 0$. Then, for every $\varepsilon > 0$ there exist $s_\varepsilon \in [0, \infty)$, $\hat{x}_\varepsilon \in H$, $\|\hat{x}_\varepsilon\| = 1$, such that $\langle Q(s_\varepsilon)\hat{x}_\varepsilon, \hat{x}_\varepsilon \rangle < \varepsilon$.

Let $z_\varepsilon(t) = z(t, s_\varepsilon; \hat{x}_\varepsilon)$, $u_\varepsilon(t) = S(t)z_\varepsilon(t)$ for all $t \geq s_\varepsilon$. We get

$$\begin{aligned} \varepsilon &> \langle Q(s_\varepsilon)\hat{x}_\varepsilon, \hat{x}_\varepsilon \rangle \geq E \int_{s_\varepsilon}^{s_\varepsilon+\tau} \langle K(t)S(t)z(t, s_\varepsilon; \hat{x}_\varepsilon), S(t)z(t, s_\varepsilon; \hat{x}_\varepsilon) \rangle dt \geq \\ &\geq \delta_0 E \int_{s_\varepsilon}^{s_\varepsilon+\tau} \|u_\varepsilon(t)\|^2 dt. \end{aligned}$$

On the other hand we have

$$\begin{aligned} \varepsilon > \langle Q(s_\varepsilon) \widehat{x}_\varepsilon, \widehat{x}_\varepsilon \rangle \geq E \int_{s_\varepsilon}^{s_\varepsilon + \tau} \|C(t) z_\varepsilon(t)\|^2 dt \text{ and} \\ \varepsilon \geq 1/2E \int_{s_\varepsilon}^{s_\varepsilon + \tau} \|C(t)y(t, s_\varepsilon; \widehat{x}_\varepsilon)\|^2 dt - \\ - \widetilde{C}^2 \int_{s_\varepsilon}^{s_\varepsilon + \tau} E \|y(t, s_\varepsilon; \widehat{x}_\varepsilon) - z(t, s_\varepsilon; \widehat{x}_\varepsilon)\|^2 dt, \end{aligned} \quad (3.9)$$

where $y(t, s_\varepsilon; \widehat{x}_\varepsilon)$ is the mild solution of $\{A; G_i\}$ with the initial condition $y(s_\varepsilon) = \widehat{x}_\varepsilon$ and $\widetilde{C} = \sup_{0 \leq r < \infty} \|C(r)\|$. We need an upper estimation for

$$\int_{s_\varepsilon}^{s_\varepsilon + \tau} E \|y(t, s_\varepsilon; \widehat{x}_\varepsilon) - z(t, s_\varepsilon; \widehat{x}_\varepsilon)\|^2 dt. \text{ From Lemma 3, (1.1) and (3.5) we get } z(t, s_\varepsilon; \widehat{x}_\varepsilon) - y(t, s_\varepsilon; \widehat{x}_\varepsilon) = \int_{s_\varepsilon}^t U(t, r)B(r)S(r)z(r, s_\varepsilon; \widehat{x}_\varepsilon)dr +$$

$$+ \sum_{i=1}^m \int_{s_\varepsilon}^t U(t, r)G_i(r)[z(r, s_\varepsilon; \widehat{x}_\varepsilon) - y(r, s_\varepsilon; \widehat{x}_\varepsilon)]dw_i(r).$$

Consider $\widetilde{B} = \sup_{0 \leq r < \infty} \|B(r)\|$. Taking mean square, using P_2 and Fubini's theorem it follows $E \|y(t, s_\varepsilon; \widehat{x}_\varepsilon) - z(t, s_\varepsilon; \widehat{x}_\varepsilon)\|^2 \leq (m+1)M_0^2 e^{2\omega\tau}[\widetilde{B}^2\varepsilon/\delta_0 +$
 $+ \sum_{i=1}^m \widetilde{G}_i^2 \int_{s_\varepsilon}^t E \|z(r, s_\varepsilon; \widehat{x}_\varepsilon) - y(r, s_\varepsilon; \widehat{x}_\varepsilon)\|^2 dr]$ for all $t \in [s_\varepsilon, s_\varepsilon + \tau]$,

because $\varepsilon \geq \delta_0 E \int_{s_\varepsilon}^{s_\varepsilon + \tau} \|u_\varepsilon(t)\|^2 dt$. By Gronwall's inequality and since $t \in [s_\varepsilon, s_\varepsilon + \tau]$ we obtain $E \|y(t, s_\varepsilon; \widehat{x}_\varepsilon) - z(t, s_\varepsilon; \widehat{x}_\varepsilon)\|^2 \leq \varepsilon r_1$, where
 $r_1 = (m+1)(M_0^2 e^{2\omega\tau} \widetilde{B}^2 1/\delta_0) \exp(\tau M_0^2 e^{2\omega\tau} (m+1) \sum_{i=1}^m \widetilde{G}_i^2)$.

Since $\{A, G_i; C\}$ is uniformly observable, (3.9) becomes $\varepsilon \geq (1/2)\gamma - r_1 \widetilde{C}^2 \varepsilon$. We get a contradiction since $\varepsilon > 0$ is arbitrary. Hence there exists $r_2 > 0$ such that

$$\langle Q(t)x, x \rangle \geq r_2 \|x\|^2 \quad (3.10)$$

for all $t \in \mathbf{R}_+$ and $x \in H$.

If $P_n(t) = P_n(s+\tau, t, P(s+\tau))$ is the classical solution of (3.3) with the initial condition $P_n(s+\tau) = P(s+\tau)$ and $P(s)$ is the bounded solution

of (3.1), then we apply the Ito's formula for the classical solution of (3.6) and the function $v_n(t, x) = \langle P_n(t)x, x \rangle$. Taking expectations, we get

$$\begin{aligned} & E \langle P_n(s + \tau)z_n(s + \tau, s; x), z_n(s + \tau, s; x) \rangle - \langle P_n(s)x, x \rangle = \\ & = -E \int_s^{s+\tau} \|C(t)z_n(t, s; x)\|^2 + \langle K(t)S_n(t)z_n(t, s; x), S_n(t)z_n(t, s; x) \rangle dt, \end{aligned}$$

where $S_n(t) = K(t)B^*(t)P_n(t)$. As $n \rightarrow \infty$ in the last equality and using Lemma 2 and 3 we get $E \langle P(s + \tau)z(s + \tau, s; x), z(s + \tau, s; x) \rangle - \langle P(s)x, x \rangle = -E \int_s^{s+\tau} \|C(t)z(t, s; x)\|^2 + \langle K(t)S(t)z(t, s; x), S(t)z(t, s; x) \rangle dt$.

Therefore we get $\langle P(s)x, x \rangle - E \langle P(s + \tau)z(s + \tau, s; x), z(s + \tau, s; x) \rangle = \langle Q(s)x, x \rangle$ for all $s \geq 0$ and $x \in H$. From (3.10) and since $P(\cdot)$ is bounded on \mathbf{R}_+ and nonnegative, we deduce that there exists $r_3 > 0$ such as

$$r_3 \|x\|^2 \geq \langle P(s)x, x \rangle \geq \langle Q(s)x, x \rangle \geq r_2 \|x\|^2 \quad (3.11)$$

for all $s \geq 0$ and $x \in H$. If we take $\delta = r_2$ it follows a).

Let us prove b). From (3.11) we get $\frac{r_2}{r_3} \langle P(s)x, x \rangle \leq r_2 \|x\|^2 \leq$

$$\leq \langle P(s)x, x \rangle - E \langle P(s + \tau)z(s + \tau, s; x), z(s + \tau, s; x) \rangle$$

and $(1 - \frac{r_2}{r_3}) \langle P(s)x, x \rangle \geq E \langle P(s + \tau)z(s + \tau, s; x), z(s + \tau, s; x) \rangle$.

Let $q = (1 - \frac{r_2}{r_3}) < 1$, $v(s, x) = \langle P(s)x, x \rangle$ and $g(s, x) = Ev(s + \tau, z(s + \tau, s; x))$, $s \in \mathbf{R}_+$, $x \in H$. Then

$$g(s, x) \leq qv(s, x) \quad (3.12)$$

for any $s \in \mathbf{R}_+$, $x \in H$. We will prove that for all $s, p \in \mathbf{R}_+$ and $u \in H$

$$Eg(s, z(s, p, u)) = Ev(s + \tau, z(s + \tau, p; x)). \quad (3.13)$$

Let $\xi \in L^s(H)$ be a simple random variable (it takes on only a finite number of values) then $\xi = \sum_{i=1}^n \chi_{A_i} x_i$, where $\{A_i\}_{i=1, \dots, m}$, $A_i \in F_s$ is a partition of Ω , χ_{A_i} is the characteristic function of A_i and $x_i \in H$. Then $g(s, \xi) = \sum_{i=1}^n \chi_{A_i} g(s, x_i)$. By the linear dependence of $z(s + \tau, s; x)$ to the initial conditions and since $z(s + \tau, s; x)$ is F_s - independent we obtain

$$\begin{aligned} & E \langle P(s + \tau)z(s + \tau, s; \xi), z(s + \tau, s; \xi) \rangle = \\ & = E \left\langle P(s + \tau)z(s + \tau, s; \sum_{i=1}^n \chi_{A_i} x_i), z(s + \tau, s; \sum_{i=1}^n \chi_{A_i} x_i) \right\rangle = \\ & = E \left(\sum_{i=1}^n \chi_{A_i} \langle P(s + \tau)z(s + \tau, s; x_i), z(s + \tau, s; x_i) \rangle \right) = \\ & = \sum_{i=1}^n P(A_i)g(s, x_i) = Eg(s, \xi). \end{aligned}$$

Consequently

$$Eg(s, \xi) = E(\langle P(s + \tau)z(s + \tau, s; \xi), z(s + \tau, s; \xi) \rangle). \quad (3.14)$$

Since $z(s, p, u) \in L^s(H)$ for all $s \geq p \geq 0$ and $u \in H$, there exists a sequence $\{\xi_n\}_{n \in N} \subset L^s(H)$ of simple random variables, which converges to $z(s, p, u)$ and satisfies (3.14).

Thus $g(s, \xi_n) = E(\langle P(s + \tau)z(s + \tau, s; \xi_n), z(s + \tau, s; \xi_n) \rangle)$.

As $n \rightarrow \infty$ in the last equality we get

$$Eg(s, z(s, p, u)) = E(\langle P(s + \tau)z(s + \tau, s; z(s, p, u)), z(s + \tau, s; z(s, p, u)) \rangle)$$

From the uniqueness of the mild solution of (3.4) we deduce

$$Eg(s, z(s, p, u)) = E(\langle P(s + \tau)z(s + \tau, p, u), z(s + \tau, p, u) \rangle).$$

We obtain (3.13). Then from (3.12) we deduce

$$E[g(s, z(s, p; x))] \leq qEv(s, z(s, p; x)) \text{ or}$$

$$Ev(s + \tau, z(s + \tau, p; x)) \leq qEv(s, z(s, p; x))$$

From the last inequality, Lemma 4 and (3.11) it follows that there exist $\beta \geq 1$ and $\alpha > 0$ ($\alpha = -\frac{\ln q}{\tau}$) such as $E \|z(t, p; x)\|^2 \leq \beta e^{-\alpha(t-p)} \|x\|^2$ for all $t \geq p \geq 0$ and $x \in H$. Hence P is a stabilizing solution. ■

The following results are the infinite dimensional versions of Proposition 5 and Theorem 1 from [7].

Proposition 1 *Under the assumptions of the above theorem the Riccati equation (3.1) has at most one nonnegative bounded solution.*

Proof. Since it is not very difficult to prove (see Corollary 3.2 in [3]) that any bounded and stabilizing solution of (3.1) is maximal (see [3]) in the class of bounded solutions, the conclusion follows from the above theorem . ■

The following theorem is the main result of this section.

Theorem 2 *Assume $\{A, G_i; B\}$ is stabilizable and $\{A, G_i; C\}$ is uniformly observable. Then the Riccati equation (3.1) has a unique nonnegative bounded on \mathbf{R}_+ solution $P(t)$, which is a stabilizing solution and there exists $\delta > 0$ such that $P(t) \geq \delta I$ for all $t \in [0, \infty)$.*

Proof. From Theorem 4.1 in [3] it follows that under stabilizability conditions the equation (3.1) has a bounded solution on \mathbf{R}_+ . From Theorem 1 and the above proposition we deduce the conclusion. ■

References

- [1] W. Grecksch, C. Tudor, Stochastic Evolution equations, A Hilbert Space Approach Math. Res. Vol 75, Acad. Verlag, 1995
- [2] A. Pazy, Semigroups of linear operators and applications to partial differential equations, Applied Mathematical Sciences 44, Springer-Verlag, Berlin, New York, 1983.

- [3] G. Da Prato, A. Ichikawa, Quadratic control for linear time-varying systems, SIAM J. Control and Optimization, vol 28, no. 2, 1990,359-381.
- [4] G. Da Prato, A. Ichikawa, Lyapunov equations for time-varying linear systems, Systems and Control Letters 9(1987) 165-172.
- [5] G. Da Prato, J. Zabczyk, Stochastic Equations in Infinite Dimensions, University Press Cambridge, 1992.
- [6] A.J. Pritchard, J. Zabczyk, Stability and stabilizability of infinite dimensional systems, SIAM Review, vol 23, no. 1, 1981.
- [7] T. Morozan, Stochastic uniform observability and Riccati equations of stochastic control, Rev. Roumaine Math. Pures Appl., 38(1993), 9, pp. 771-481.
- [8] T. Morozan, On the Riccati equation of stochastic control, International Series on Numerical Mathematics, vol. 107, Birkhauser Verlag Basel, 1992.
- [9] V. Ungureanu, "A Counter example for the implication "uniform controllability implies stabilizability" in the stochastic case" ,Seminar on Mathematical Analysis and Applications in control Theory", Timisoara, no. 118, 2000

COMPONENTWISE ASYMPTOTIC STABILITY INDUCED BY SYMMETRICAL POLYHEDRAL TIME-DEPENDENT CONSTRAINTS

Mihail Voicu and Octavian Pastravanu

Department of Automatic Control and Industrial Informatics

Technical University "Gh. Asachi" of Iasi

Bvd. Mangeron 53A, 6600 Iasi, Romania

mvoicu@ac.tuiasi.ro

Abstract In this paper the concepts of componentwise asymptotic stability with respect to a differentiable vector function $\mathbf{h}(t)$ (approaching 0 as $t \rightarrow \infty$) (CWAS_h) and componentwise exponentially asymptotic stability (CWEAS), previously introduced, have been extended to Q - CWAS_h and Q - CWEAS (Q being a $q \times n$ real matrix), respectively, in order to cover the more general situation of polyhedral time-dependent flow-invariant sets, defined by $|Q\mathbf{x}| \leq \mathbf{h}(t)$, $\mathbf{x} \in \mathbb{R}^n$, $t \in \mathbb{R}_+$, symmetrical with respect to the equilibrium point of a given continuous-time linear system $\dot{\mathbf{x}} = Q\mathbf{x}$, $t \in \mathbb{R}_+$, $\mathbf{x} \in \mathbb{R}^n$. It is proved that Q - CWAS_h is equivalent with the existence of a $q \times q$ matrix \mathbf{E} such that $\mathbf{EQ} = Q\mathbf{A}$, $\bar{\mathbf{E}}\mathbf{h}(t) \leq \mathbf{h}(t)$, where the bar operator ($\bar{}$) transforms only the extra diagonal elements of \mathbf{E} into their corresponding absolute values and does not change its diagonal elements. By specializing vector function $\mathbf{h}(t)$ in an exponentially decaying form, the concept of Q - CWEAS is characterized by the above mentioned matrix equation and an algebraic inequality. For $Q = \mathbf{I}_n$ these results consistently yield the earlier ones. As in this case, there exists a strong connection between Q - CWAS_h (Q - CWEAS) and the asymptotic stability, but now this connection is amended by the observability of the pair (Q, A) .

Keywords: Stability analysis, Flow-invariant sets, Time-invariant linear systems, Continuous-time systems, Discrete-time systems

1. Introduction

Consider the nonlinear dynamical system described by the differential equation:

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}), \quad t \in \mathbb{R}_+, \quad \mathbf{x} \in \mathbb{R}^n, \quad (1)$$

with

$$\mathbf{f}(t, 0) = 0, \quad t \in \mathbb{R}_+, \quad (2)$$

and the initial condition:

$$\mathbf{x}(t_0) = \mathbf{x}_0, \quad t_0 \in \mathbb{R}_+, \quad (3)$$

where \mathbf{f} ensures the existence and the uniqueness of the Cauchy solution on the time interval $[t_0, +\infty)$.

The purpose of this paper is to extend the concepts of componentwise asymptotic stability (CWAS) and of componentwise exponential asymptotic stability (CWEAS), defined and characterized in previous works ([15], [16], [17]), to polyhedral flow-invariant sets, symmetrical with respect to the equilibrium point of system (1) (according to (2)):

$$\mathbf{x} = 0. \quad (4)$$

In order to consistently specify the extensions taken into consideration, let us remind first some notations already used in the above mentioned works. Let $\mathbf{v} =: (v_i)$ and $\mathbf{w} =: (w_i)$ be two vectors of the same dimension. We denote by $|\mathbf{v}|$ the vector with the components $|v_i|$ and by $\mathbf{v} \leq \mathbf{w}$ or by $\mathbf{v} > \mathbf{w}$ the componentwise inequalities $v_i \leq w_i$ or $v_i > w_i$ respectively.

Given a matrix $\mathbf{Q} \in \mathbb{R}^{q \times n}$, with $\text{rank } \mathbf{Q} = \min(n, q) \geq 1$, and a continuous differentiable vector function:

$$\mathbf{h} : \mathbb{R}_+ \rightarrow \mathbb{R}^q, \quad (5)$$

assume that the following conditions hold:

$$\mathbf{h}(t) > 0, \quad t \in \mathbb{R}_+, \quad (6)$$

$$\lim_{t \rightarrow \infty} \mathbf{h}(t) = 0. \quad (7)$$

The envisaged extensions refer to the following two definitions.

Definition 1 The system (1) is called *Q-componentwise asymptotically stable with respect to $\mathbf{h}(t)$* ($Q\text{-CWAS}_h$), if for each $t_0 \in \mathbb{R}_+$ and for each \mathbf{x}_0 with

$$|\mathbf{Q}\mathbf{x}_0| \leq \mathbf{h}(t_0), \quad (8)$$

the Cauchy solution of (1) satisfies:

$$|\mathbf{Q}\mathbf{x}(t)| \leq \mathbf{h}(t) \quad \text{for each } t \geq t_0. \quad (9)$$

Definition 2 The system (1) is called *Q-componentwise exponential asymptotically stable* ($Q\text{-CWEAS}$) if there exist a positive vector $\mathbf{d} > 0$, $\mathbf{d} \in \mathbb{R}^q$, and a negative scalar $r < 0$ such that system (1) is $Q\text{-CWAS}_h$ for:

$$\mathbf{h}(t) = \mathbf{d}e^{rt}. \quad (10)$$

The characterization of $Q\text{-CWAS}_h$ and $Q\text{-CWEAS}$ will be performed by using the flow-invariance method for which the following basic result is available ([6]).

Theorem 1 A time-dependent compact set $X(t) \subset \mathbb{R}^n$, $t \in \mathbb{R}_+$, is flow-invariant for system (1) (i.e. for each $t_0 \in \mathbb{R}_+$ and for each $\mathbf{x}_0 \in X(t_0)$ the solution of (1), (2) satisfies $\mathbf{x}(t) \in X(t)$ for each $t \geq t_0$) if and only if

$$\lim_{\tau \downarrow 0} \tau^{-1} \text{dist}(\mathbf{v} + \tau \mathbf{f}(t, \mathbf{v}); X(t + \tau)) = 0 \quad (11)$$

for each $t \in \mathbb{R}_+$ and for each $\mathbf{v} \in X(t)$.

In relation (11) $\text{dist}(\mathbf{v}; X) = \inf_{\mathbf{w} \in X} \text{dist}(\mathbf{v}; \mathbf{w})$ denotes the distance from $\mathbf{v} \in \mathbb{R}^n$ to the set X .

The concept of flow-invariant time-dependent sets has been exploited in several works for studying particular properties of the solutions of various types of differential equations and is based on the pioneering researches in ([5], [2], [3], [4]). A remarkable monograph on this field is due to Pavel ([6]). The use of time-dependent rectangular sets $X(t) \subset \mathbb{R}^n$, $t \in \mathbb{R}_+$, has been proposed by ([15], [16]) for continuous-time linear constant systems, resulting in the definition and analysis of special types of stability, namely the CWAS_h and CWEAS. An overview on the application of the flow-invariance method in control theory and design is presented in ([17]), including the case of continuous-time nonlinear dynamical systems. Exploiting the inequality-form of the characterizations generated by time-dependent rectangular sets $X(t) \subset \mathbb{R}^n$, $t \in \mathbb{R}_+$, further results on linear interval matrix systems, disturbed systems, uncertain systems, and a class of nonlinear systems have been reported in ([7], [8], [9], [10], [11], [12], [13], [14]).

In order to characterize the concepts of Q-CWAS_h and Q-CWEAS (Definitions 1 and 2) by using Theorem 1, the following special type of time-dependent polyhedral set will be considered as flow-invariant set:

$$X(t) =: \{\mathbf{v} \in \mathbb{R}^n; |\mathbf{Q}\mathbf{v}| \leq \mathbf{h}(t)\} \subset \mathbb{R}^n, \quad t \in \mathbb{R}_+. \quad (12)$$

Remark 1 Under these circumstances it is obvious that, by taken $q = n$ in (5) and $\mathbf{Q} = \mathbf{I}_n$ (the unit matrix of order n) in (8) and (9) (but originally in (12)), the Definitions 1 and 2 consistently yield the previously introduced concepts of CWAS_h and CWEAS respectively, defined and characterized in ([15], [16], [17]) and ([7], [8], [9], [10], [11], [12], [13], [14]). In the case of an arbitrary $\mathbf{Q} \in \mathbb{R}^{n \times n}$, with rank $\mathbf{Q} = n$, Q-CWAS_h and CWEAS operate in $\text{Im } \mathbf{Q} = \mathbb{R}^n$ but in a new vector basis of \mathbb{R}^n , different from that one in which system (1) is initially expressed. As mentioned in ([17]), by using the state similarity transformation $\tilde{\mathbf{x}} = \mathbf{Q}\mathbf{x}$ for system (1), the special CWAS_h and CWEAS and their characterizations for the corresponding transformed system are to be approached. ■

In Section 2 the characterizations in view of Definitions 1 and 2 only for the linear constant continuous-time dynamical systems are per-

formed. The concluding remarks and several comments related to some already existing particular results – see the survey paper ([1]) and the papers cited therein – are included in Section 3.

2. Linear constant continuous-time dynamical systems

System (1) is described by the following differential equation:

$$\dot{\mathbf{x}} = \mathbf{Ax}, \quad t \in \mathbb{R}_+, \quad \mathbf{x} \in \mathbb{R}^n. \quad (13)$$

Remark 2 In terms of the definition of flow-invariant set $X(t)$ (given by (12) with (5) – (7)), which is equivalent to Definition 1, the following state transformation is considered:

$$\mathbf{y} = \mathbf{Qx}, \quad \mathbf{x} \in \mathbb{R}^n, \quad \mathbf{y} \in \mathbb{R}^q, \quad (14)$$

and, corresponding to system (13), the possible existence of the transformed system has to be taken into account:

$$\dot{\mathbf{y}} = \mathbf{Ey}, \quad t \in \mathbb{R}_+, \quad \mathbf{y} \in \mathbb{R}^q. \quad (15)$$

Actually system (15) may represent system (13) in $\text{Im } \mathbf{Q}$. For this purpose there exists a system matrix $\mathbf{E} \in \mathbb{R}^{q \times q}$ if and only if the following consistency condition holds:

$$\mathbf{QA}(\mathbf{Q}^T \mathbf{Q} - \mathbf{I}_n) = 0, \quad (16)$$

and \mathbf{E} can be calculated with:

$$\mathbf{E} = \mathbf{QAQ}^T + \hat{\mathbf{E}}(\mathbf{QQ}^T - \mathbf{I}_q), \quad (17)$$

where $\hat{\mathbf{E}}$ is an arbitrary matrix of order q , \mathbf{I}_q is the unit matrix of the same order, and \mathbf{Q}^T is an inverse of \mathbf{Q} , namely (according to the case): it is a right inverse \mathbf{Q}^R , $\mathbf{QQ}^R = \mathbf{I}_q$ (for $q < n$), the regular inverse \mathbf{Q}^{-1} (for $q = n$) or a left inverse \mathbf{Q}^L , $\mathbf{Q}^L\mathbf{Q} = \mathbf{I}_n$ (for $q > n$).

It is a simple matter to see that:

(i) in the case $q < n$ there exists a matrix \mathbf{E} if and only if (16) is satisfied with $\mathbf{Q}^T = \mathbf{Q}^R$, i.e. the following consistency condition is met:

$$\mathbf{QA}(\mathbf{Q}^R \mathbf{Q} - \mathbf{I}_n) = 0; \quad (18)$$

(ii) in the other case, $q \geq n$, always exists a matrix \mathbf{E} given by (17) because (16) is satisfied for any $\mathbf{Q}^T = \mathbf{Q}^L$. ■

For the concise writing of the next result, let us remind that for a given square real matrix $\mathbf{M} =: (m_{ij})$ we denote by $\bar{\mathbf{M}} =: (\bar{m}_{ij})$ the matrix with $\bar{m}_{ii} = m_{ii}$ and $\bar{m}_{ij} = |m_{ij}|$, $i \neq j$.

Theorem 2 System (13) is Q-CWAS_h if and only if there exists a matrix $\mathbf{E} \in \mathbb{R}^{q \times q}$ such that:

$$\mathbf{EQ} = \mathbf{QA}, \quad (19)$$

$$\bar{\mathbf{E}}\mathbf{h}(t) \leq \dot{\mathbf{h}}(t), \quad t \in \mathbb{R}_+. \quad (20)$$

Proof. The Q-CWAS_h of system (13), i.e. the flow-invariance of the set $X(t)$ (given by (12) with (5) – (7)) is equivalent to the following two conditions:

- (i) on the one hand (according to Remark 2), the existence of system (15), i.e. of matrix \mathbf{E} given by (19) which is expressed by (17) (either for any $q \geq n$, or for any $q < n$ if and only if (18) holds);
- (ii) on the other hand (according to Theorem 1), the inequality:

$$|\mathbf{Q}[\mathbf{v} + \tau(\mathbf{Av} + \mathbf{w}(\tau))]| \leq \mathbf{h}(t + \tau), \quad (21)$$

which must be componentwise fulfilled for each $t \in \mathbb{R}_+$, for each \mathbf{v} with $|\mathbf{Q}\mathbf{v}| \leq \mathbf{h}(t)$, for $\tau > 0$, small enough, and for a certain $\mathbf{w} : \mathbb{R}_+ \rightarrow \mathbb{R}^n$, with $\mathbf{w}(\tau) \rightarrow 0$ as $\tau \downarrow 0$.

Now, combining (19) and (21) it equivalently results:

$$|\mathbf{Q}\mathbf{v} + \tau(\mathbf{EQ}\mathbf{v} + \mathbf{Q}\mathbf{w}(\tau))| \leq \mathbf{h}(t + \tau), \quad (22)$$

which must be componentwise fulfilled for each $t \in \mathbb{R}_+$, for each \mathbf{v} with $|\mathbf{Q}\mathbf{v}| \leq \mathbf{h}(t)$, for $\tau > 0$, small enough, and for $\mathbf{w}(\tau) \rightarrow 0$ as $\tau \downarrow 0$.

According to the differentiability of $\mathbf{h}(t)$ there exists $\mathbf{z} : \mathbb{R}_+ \rightarrow \mathbb{R}^n$, with $\mathbf{z}(\tau) \rightarrow 0$ as $\tau \downarrow 0$, such that $\mathbf{h}(t+\tau) - \mathbf{h}(t) = \tau \dot{\mathbf{h}}(t) + \tau \mathbf{z}(\tau)$, $t \in \mathbb{R}_+$.

Thus, (22) is equivalent to

$$|\mathbf{Q}\mathbf{v} + \tau(\mathbf{EQ}\mathbf{v} + \mathbf{Q}\mathbf{w}(\tau))| \leq \mathbf{h}(t) + \tau \dot{\mathbf{h}}(t) + \tau \mathbf{z}(\tau), \quad (23)$$

which must be componentwise fulfilled for each $t \in \mathbb{R}_+$, for each \mathbf{v} with $|\mathbf{Q}\mathbf{v}| \leq \mathbf{h}(t)$, for $\tau > 0$, small enough, for $\mathbf{w}(\tau) \rightarrow 0$ as $\tau \downarrow 0$, and for $\mathbf{z}(\tau) \rightarrow 0$ as $\tau \downarrow 0$.

Using transformation (14), rewritten as $\mathbf{u} = \mathbf{Q}\mathbf{v}$, it follows that the vectorial inequality (23) is equivalent to:

$$|\mathbf{u} + \tau(\mathbf{Eu} + \mathbf{Q}\mathbf{w}(\tau))| \leq \mathbf{h}(t) + \tau \dot{\mathbf{h}}(t) + \tau \mathbf{z}(\tau) \quad (24)$$

and this must componentwise hold for each $t \in \mathbb{R}_+$, for each $\mathbf{u} \in Im\mathbf{Q}$ with $|\mathbf{u}| \leq \mathbf{h}(t)$, for $\tau > 0$, small enough, for $\mathbf{w}(\tau) \rightarrow 0$ as $\tau \downarrow 0$, and for $\mathbf{z}(\tau) \rightarrow 0$ as $\tau \downarrow 0$.

It is obvious that (24) must also hold for the maximum value and for the minimum value of each component of $\mathbf{u} + \tau \mathbf{Eu}$ for $\tau > 0$, small enough, for $t \in \mathbb{R}_+$ and for each $\mathbf{u} \in Im\mathbf{Q}$ with $|\mathbf{u}| \leq \mathbf{h}(t)$. Since $\mathbf{u} + \tau \mathbf{Eu}$ is linear for \mathbf{u} and the set $X(t)$, given by (12) and rewritten as:

$$X(t) =: \{\mathbf{u} \in Im \mathbf{Q}; |\mathbf{u}| \leq \mathbf{h}(t)\}, \quad t \in \mathbb{R}_+, \quad (25)$$

is symmetrical with respect to $\mathbf{x} = 0$, the extrema of the i -th component of $\mathbf{u} + \tau \mathbf{E}\mathbf{u}$ for $\tau > 0$, small enough, can be reached, respectively, for

$$\mathbf{u}_{ex}^i = \pm \text{diag}\{\text{sgn}e_{i1}, \dots, \text{sgn}e_{i(i-1)}, 1, \text{sgn}e_{ii+1}, \dots, \text{sgn}e_{iq}\} \mathbf{h}(t) \in X(t), \quad i=1, 2, \dots, q, \quad (26)$$

where e_{ij} , $i, j = 1, 2, \dots, q$, are the elements of matrix \mathbf{E} . Thus, for $\mathbf{u} = \mathbf{u}_{ex}^i$ the i -th inequality from (24), after simplification by $\tau > 0$, is equivalent to:

$$e_{ii}h_i(t) + \sum_{j=1, j \neq i}^q |e_{ij}|h_j(t) \leq \dot{h}_i(t) + z_i(\tau) \mp w_i(\tau), \quad i=1, 2, \dots, q, \quad (27)$$

for each $t \in \mathbb{R}_+$, for $\tau > 0$, small enough, for $\mathbf{w}(\tau) \rightarrow 0$ and for $\mathbf{z}(\tau) \rightarrow 0$ as $\tau \downarrow 0$, where $h_i(t)$, $z_i(\tau)$ and $w_i(\tau)$ are the components of $\mathbf{h}(t)$, $\mathbf{z}(\tau)$ and $\mathbf{w}(\tau)$, respectively.

Now, taking into account that $\mathbf{z}(\tau) \rightarrow 0$ and $\mathbf{w}(\tau) \rightarrow 0$ as $\tau \downarrow 0$, the equivalence between (24) and (15) is proved. ■

To this extent it is obvious that, according to Theorems 2 and 3 in ([16]), the following results can be stated.

Theorem 3 System (1) is Q -CWAS _{h} if and only if there exists a matrix $\mathbf{E} \in \mathbb{R}^{q \times q}$ such that (19) and the following inequality are met:

$$e^{\bar{\mathbf{E}}(t-\vartheta)} \mathbf{h}(\vartheta) \leq \mathbf{h}(t), \quad t \geq \vartheta \geq 0. \quad (28)$$

Theorem 4 A necessary and sufficient condition for the existence of $\mathbf{h}(t)$ such that system (1) be Q -CWAS _{h} is the existence of matrix $\mathbf{E} \in \mathbb{R}^{q \times q}$ satisfying (19) and $\bar{\mathbf{E}}$ be Hurwitzian.

Remark 3 Let \mathcal{H} be the Abelian semigroup of the solutions of (20) in the conditions of Theorem 4. Obviously, system (13) is Q -CWAS _{h} for each $\mathbf{h} \in \mathcal{H}$. Moreover, for each pair \mathbf{h}_1 and \mathbf{h}_2 the Q -CWAS _{h_1} is equivalent to CWAS _{h_2} . This allows us to specialize $\mathbf{h}(t)$ and to characterize in a more explicit manner the free response of system (13), namely according to Definition 2. ■

Theorem 5 System (1) is Q -CWEAS if and only if there exists a matrix $\mathbf{E} \in \mathbb{R}^{q \times q}$ such that (19) and the following inequality are met:

$$\bar{\mathbf{E}}\mathbf{d} \leq_r \mathbf{d}. \quad (29)$$

Proof. It is immediate by replacing (10) into (20). ■

In view of Remark 3 the following statement is obvious.

Theorem 6 System (13) is Q -CWAS _{h} if and only if it is Q -CWEAS.

In the light of these results and according to Theorem 4 in ([7]) it is quite natural to state next the conditions that ensure the compatibility of inequality (29) regardless of its meaning in connection with CWEAS of system (13). For this purpose let us denote by $\lambda_i(\bar{\mathbf{E}})$, $i = 1, \dots, q$, the eigenvalues of $\bar{\mathbf{E}}$.

Theorem 7 a. $\bar{\mathbf{E}}$ has a real eigenvalue (simple or multiple), denoted by $\lambda_{\max}(\bar{\mathbf{E}})$, which fulfils the dominance condition:

$$\operatorname{Re}[\lambda_i(\bar{\mathbf{E}})] \leq \lambda_{\max}(\bar{\mathbf{E}}), \quad i = 1, \dots, q. \quad (30)$$

b. Inequality (29) is compatible if and only if

$$\lambda_{\max}(\bar{\mathbf{E}}) \leq r. \quad (31)$$

Now, according to Theorem 8 in ([17]), the Q-CWEAS of system (13) can be further characterized as follows.

Theorem 8 System (13) is Q-CWEAS if and only if there exists a matrix $\mathbf{E} \in \mathbb{R}^{q \times q}$ such that (19) and one of the following equivalent conditions are met:

$$(i) \quad (-1)^k \bar{E}_k > 0, \quad k = 1, \dots, q, \quad (32)$$

where \bar{E}_k , $k = 1, \dots, q$, are the leading principal minors of $\bar{\mathbf{E}}$;

$$(ii) \quad \det \bar{\mathbf{E}} \neq 0, \quad (-\bar{\mathbf{E}})^{-1} \geq 0, \quad (33)$$

where the inequality is to be taken elementwise;

$$(iii) \quad \bigcup_{i=1}^q G_i(\mathbf{E}_d) \subset \{s \in \mathbb{C}; \operatorname{Res} < 0\}, \quad (34)$$

where $G_i(\mathbf{E}_d) = \{s \in \mathbb{C}; |s - e_{ii}| \leq d_i^{-1} \sum_{j=1, j \neq i}^q |e_{ij}| d_j\}$, $i = 1, \dots, q$, are the Gershgorin's discs associated to matrix $\mathbf{E}_d = \operatorname{diag}\{d_1^{-1}, \dots, d_q^{-1}\} \times \mathbf{E} \operatorname{diag}\{d_1, \dots, d_q\}$,

i.e. to \mathbf{E} and vector \mathbf{d} (having the components d_1, \dots, d_n);

$$(iv) \quad \lambda_{\max}(\bar{\mathbf{E}}) \leq r < 0. \quad (35)$$

Unlike the special forms of CWAS_h and CWEAS, i.e. Q-CWAS_h and Q-CWEAS for $\mathbf{Q} = \mathbf{I}_n$, which are sufficient conditions for the asymptotic stability of systems (13) because $\mathbf{E} = \mathbf{A}$, in the case of an arbitrary \mathbf{Q} according to (19), the relation between Q-CWAS_h (Q-CWEAS) and the asymptotic stability depends on the pair (\mathbf{Q}, \mathbf{A}) . Obviously, in the context of Q-CWAS_h (Q-CWEAS), the dynamics of system (13) is actually observed by means of the transformation (14). As a matter of fact the state observability of system (13), (14), i.e. of the pair (\mathbf{Q}, \mathbf{A}) , plays here an adequate part and the following results will clarify its place in the mentioned relation.

Theorem 9 *The observability properties of the pair (\mathbf{Q}, \mathbf{A}) are as follows:*

- For $q < n$, system (13),(14) (pair (\mathbf{Q}, \mathbf{A})) is partially state observable; the dimension of the completely observable part of system (13),(14) is q .
- For $q \geq n$, system (13), (14) (pair (\mathbf{Q}, \mathbf{A})) is completely state observable.

Proof. It relies on the rank evaluation of (\mathbf{Q}, \mathbf{A}) - observability matrix performed as follows:

$$\begin{aligned} \text{rank } \mathbf{O} &= \text{rank} \begin{bmatrix} \mathbf{Q} \\ \mathbf{Q}\mathbf{A} \\ \vdots \\ \mathbf{Q}\mathbf{A}^{n-1} \end{bmatrix} = \text{rank} \begin{bmatrix} \mathbf{Q} \\ \mathbf{E}\mathbf{Q} \\ \vdots \\ \mathbf{E}^{q-1}\mathbf{Q} \end{bmatrix} \\ &= \text{rank} \begin{bmatrix} \mathbf{I}_q \\ \mathbf{E} \\ \vdots \\ \mathbf{E}^{q-1} \end{bmatrix} \mathbf{Q} = \min(n, q). \end{aligned} \quad (36)$$

The second equality in (3) becomes obvious by using (19) repeatedly. Further we take into consideration that:

for $q < n$, $\text{rank } \mathbf{Q} = \min(n, q) = q$ and $\text{rank } \mathbf{O} = q$;

for $q \geq n$, $\text{rank } \mathbf{Q} = \min(n, q) = n$, and $\text{rank } \mathbf{O} = n$. ■

In view of this result we can naturally state the next one.

Theorem 10 *System (13) is asymptotically stable if it is Q -CWAS_h (Q -CWEAS) and one of the following condition holds:*

- (i) $q < n$ and the unobservable part of system (13), (14), which evolves in the subspace $\text{Ker } \mathbf{O} \subset \mathbb{R}^n$ of dimension $n - q$, is asymptotically stable;
- (ii) $q \geq n$.

It is eminently clear that condition (i) reveals that, in the case $q < n$, Q -CWAS_h (Q -CWEAS) evaluates only that part of dimension q of the dynamics of system (13),(14), which is completely state observable. In this respect the following sufficient condition for the *partial asymptotic stability*, ([18]), of system (13) in the subspace $\mathbb{R}^n \setminus \text{Ker } \mathbf{O}$ may be stated too.

Theorem 11 *If system (13) is Q -CWAS_h (Q -CWEAS) and $q < n$, then the completely state observable part of system (13), (14), which evolves in the subspace $\mathbb{R}^n \setminus \text{Ker } \mathbf{O}$ of dimension q , is asymptotically stable.*

3. Concluding remarks

In this paper, the concepts of CWAS_h and CWEAS, previously introduced by the first author, have been extended to Q-CWAS_h (Definition 1) and Q-CWEAS (Definition 2), respectively, in order to cover the more general situation of polyhedral time-dependent flow-invariant sets, symmetrical with respect to the equilibrium point of a given continuous-time linear system. The main results are formulated by Theorems 2–4, where the asymptotic behavior to the infinity of such a polyhedral set is expressed by a priori defined vector function $\mathbf{h}(t)$. These novel results are consistent with those mentioned above, in the sense that the characterization of Q-CWAS_h relies on matrix operator $\bar{\mathbf{E}}$ involved in differential inequality (20), that is accompanied by matrix equation (19) generated by state-space transform (14). Note that the bar operator ($\bar{\cdot}$) is now applied to the transformed matrix \mathbf{E} , resulting from the original system matrix \mathbf{A} . Obviously, for the particularization of matrix \mathbf{Q} to the identity matrix in equation (19), Q-CWAS_h becomes CWAS_h.

By specializing vector function $\mathbf{h}(t)$ in an exponentially decaying form, the concept of Q-CWEAS is characterized in Theorem 5, which, for the studied linear case, is shown to be equivalent with Q-CWAS_h (Theorem 6). The characterization of Q-CWEAS stated in Theorem 5 is given by matrix equation (19) and algebraic inequality (29), and for the compatibility of the latter (Theorem 7) an earlier result of the authors is used. Other algebraic (and, to some extent, parametric) characterizations of Q-CWEAS are given in Theorem 8.

As in the case of CWAS_h (CWEAS), there exists a strong connection between Q-CWAS_h (Q-CWEAS) and the asymptotic stability, which is explored by Theorems 10 – 11 where, for $q < n$, the observability of the pair (\mathbf{Q}, \mathbf{A}) (Theorem 9) plays an adequate part.

Finally, it is worth mentioning that the flow invariance of the polyhedral sets with respect to linear dynamic systems has been investigated by many other works as shown in the survey paper ([1]). Unlike our approach where the time-dependence of the polyhedral sets is a priori defined, these works cover only a very special case of time-dependent polyhedral sets, when they are a posteriori proved contractive via an adequate Lyapunov function. This contractiveness is only of exponential type and, therefore, relations to Q-CWEAS are straightforward. However the framework created by the mentioned works did not address concepts related to Q-CWAS_h, just because the time-dependence of the polyhedral sets is a posteriori investigated.

References

- [1] Blanchini, F. (1999), Set invariance in control; survey paper. *Automatica*, **35**, 1747-1767.
- [2] Brezis, H. (1970), On a characterization of flow-invariant sets. *Comm. Pure Appl. Math.*, **23**, 261–263.
- [3] Crandall, M.G. (1972), A generalization of Peano's existence theorem and flow invariance. *Proc. Amer. Math. Soc.*, **36**, 151–155.
- [4] Martin, R.H. jr. (1973), Differential equations on closed subsets of a Banach space. *Trans. Amer. Math. Soc.*, **179**, 399–414.
- [5] Nagumo, M. (1942), Über die Lage der Integralkurven gewöhnlicher Differentialgleichungen. *Proc. Phys. Math. Soc. Japan*, **24**, 551–559.
- [6] Pavel, H.N. (1984), *Differential Equations Flow Invariance and Applications*. Pitman, Boston, London, Melbourne.
- [7] Păstrăvanu, O., Voicu, M. (1999), Flow-invariant rectangular sets and componentwise asymptotic stability of interval matrix systems. 5th European Control Conference, Karlsruhe, August 31 - Sept. 3, 1999; Proceedings on CD, *rubicon-Agentur für digitale Medien*, Aachen (Germany), 16 p.
- [8] Păstrăvanu, O., Voicu, M. (2000), Robustness analysis of componentwise asymptotic stability. 16th IMACS World Congress, Lausanne, August 21-25, 2000; Proceedings on CD, *©imacs* (ISBN 3 9522075 1 9), 16 p.
- [9] Păstrăvanu, O., Voicu, M. (2000), Preserving componentwise asymptotic stability under disturbances. *Revue Roumaine des Sciences Techniques* (Académie Roumaine), ser. electr. energ., t. 45, 3, 2000, pp. 413-425.
- [10] Păstrăvanu, O., Voicu, M. (2001), Dynamics of a class of nonlinear systems under flow-invariance constraints. 9th IEEE Mediterranean Conf. on Control and Automation (MED'01), Dubrovnik, June 27-29, 2001; Proceedings on CD-ROM (ISBN 953 6037 35 1), Book of abstracts (ISBN 953 6037 34 3), *IEEE*, ©KoREMA (Zagreb), 6 p.
- [11] Păstrăvanu, O., Voicu, M. (2001), Componentwise asymptotic stability of a class of nonlinear systems. 1st IFAC Symposium on System Structure and Control (SSSC 01), August 29-31, 2001, *Czech Technical University of Prague*; Book of Abstracts, p. 28; Proceedings, *IFAC*, CD-ROM, 078, 6 p.
- [12] Păstrăvanu, O., Voicu, M. (2001), Flow-invariance in exploring stability for a class of nonlinear uncertain systems. 6th European Control Conference, Porto, September 4-7, 2001, Proceedings, CD-ROM, 6 p.
- [13] Păstrăvanu, O., Voicu, M. (2001), Robustness of componentwise asymptotic stability for a class of nonlinear systems. *Proceedings of the Romanian Academy*, ser. A, vol 1, 1-2, 2001, 61-67.
- [14] Păstrăvanu, O., Voicu, M. (2002), Componentwise asymptotic stability of interval matrix systems. *Journal of Differential and Integral Equation*, vol. 15, No. 11, 1377-1394.
- [15] Voicu, M. (1984), Free response characterization via flow invariance. 9th *World Congress of IFAC*, Budapest, Preprints, **5**, 12–17.
- [16] Voicu, M. (1984), Componentwise asymptotic stability of linear constant dynamical systems. *IEEE Trans. on Aut. Control*, **10**, 937–939.
- [17] Voicu, M. (1987), On the application of the flow-invariance method in control theory and design. 10th *World Congress of IFAC*, Munich, Preprints, **8**, 364–369.
- [18] Vorotnikov, I.V., (2002), Partial stability, stabilization and control: a some recent results. 15th *World Congress of IFAC*, Barcelona, Preprints on CD-ROM, 12p.