ACADEMY OF ROMANIAN SCIENTISTS

A N N A L S

SERIES ON MATHEMATICS AND ITS APPLICATIONS

VOLUME 7 2015 NUMBER 1

ISSN 2066 – 6594

TOPICS:

- **ORDINARY AND PARTIAL DIFFERENTIAL EQUATIONS**
- **OPTIMIZATION, OPTIMAL CONTROL AND DESIGN**
- **NUMERICAL ANALYSIS AND SCIENTIFIC COMPUTING**
- **ALGEBRAIC, TOPOLOGICAL AND DIFFERENTIAL STRUCTURES**
- **PROBABILITY AND STATISTICS**
- **ALGEBRAIC AND DIFFERENTIAL GEOMETRY**
- **MATHEMATICAL MODELLING IN MECHANICS ENGINEERING SCIENCES**
- **MATHEMATICAL ECONOMY AND GAME THEORY**
- **MATHEMATICAL PHYSICS AND APPLICATIONS**

GUEST EDITOR: PEKKA NEITTAANMÄKI

EDITURA
ACADEMIEI OAMENILOR DE ȘTIINȚĂ DIN ROMÂNIA

Annals of the Academy of Romanian Scientists

Series on Mathematics and its Applications

Ann. Acad. Rom. Sci.
Ser. Math. Appl.
ISSN 2066-6594 Vol. 7, No. 1/2015

CONTENTS

Ann. Acad. Rom. Sci.
Ser. Math. Appl.
ISSN 2066-6594 Vol. 7, No. 1/2015

In memoriam

Viorel Arnăutu was my undergraduate and PHD student and also coworker in a field where he had a pioneering contribution in Romania: numerical analysis of optimal control problems governed by partial differential equations. His contribution to the development of this research direction is indeed remarkable. Computer scientist by training, he became our main specialist in scientific computation of infinite dimensional optimization problems and his premature disparition represents a big loss for our mathematical community. Viorel had a great intelect and a nice personality and will remain forever in our memory.

Viorel Barbu

Ann. Acad. Rom. Sci.
Ser. Math. Appl.
ISSN 2066-6594 Vol. 7, No. 1/2015

Viorel Arnăutu

(13.10.1955 – 04.01.2014)

Viorel Arnăutu was born on October 13-th, 1955, in Iaşi, the old capital of Moldavia, in a well-known family of physicians and professors. His whole life and career was dedicated to mathematics, that he studied in the schools and in the Department of Mathematics of the "Al. I. Cuza" University in Iaşi. He obtained his Ph.D. in Applied Mathematics in 1987 under the supervision of Academician Viorel Barbu. His scientific education also includes a postdoc grant in 1991 at the Laboratory of Numerical Analysis, CNRS and University Paris VI. After a short period at the Computer Center of the "Al. I. Cuza" University in Iaşi, his professional career included positions of assistant professor (1982 – 1992), associate professor (1992 – 2002) and full professor until his untimely death in the beginning of 2014. All the positions were in Numerical Analysis and Information Technology, at the Faculty of Mathematics, "Al. I. Cuza" University in Iaşi, Romania. He also held temporary positions for various time periods at the Institute for Applied Mathematics, Freiburg, Germany (1984) and at the Department of Mathematics, Universita degli Studi di Bari, Italy (1985). With the University of Jyväskyla (Pekka Neittaanmäki) and the Weierstrass Institute Berlin (Juergen Sprekels, Dietmar Hoemberg) a very long cooperation was active for many years, concretized in several research projects and many coauthored scientific works. Other important collaborators of Viorel Arnutu were Vincenzo Capasso, Viorel Barbu, Dan Tiba, Sebastian Aniţa. The research interests of Viorel Arnăutu span a large range of subjects: numerical methods (theory, algorithms and computer programs) for optimal control problems governed by PDE's and by variational inequalities; numerical methods for Hamilton-Jacobi equations (the synthesis of optimal control); numerical methods for PDE's and integral equations; applied mathematics, applied optimal control problems; free boundary problems: epidemic models; optimization of plates; laser hardening of steel; 3D curved mechanical

structures; population dynamics. Viorel Arnăutu is the author or co-author of four books and more than 25 papers published in international journals and proceedings volumes:

- *Optimal Control from Theory to Computer Programs* (in cooperation with Pekka Neittaanmäki), Kluwer Academic Publishers, Dordrecht, Boston, London, 2003.
- *Numerical Methods for Variational Problems*, University of Jyväskyla, Department of Mathematical Information Technology, Lecture Notes 8/2001, Jyväskyla, Finland, 2001.
- *Metode numerice pentru probleme variaţionale. Teorie şi algoritmi*, Editura Universitaţii "Alexandru Ioan Cuza", Iaşi, 2001.
- *An Introduction to Optimal Control Problems in Life Sciences and Economics. From Mathematical Models to Numerical Simulation with MatLab* (in cooperation with S. Aniţa, V. Capasso), Birkhäuser, Boston, 2010.

He had four Ph. D. students. He is survived by his wife Marcela and his two children. All his relatives, his friends and collaborators, his students will remember his very warm and pleasant personality, his positive and friendly attitude, his humour and his kindness.

Pekka Neittaanmäki

Ann. Acad. Rom. Sci.
Ser. Math. Appl.
ISSN 2066-6594 Vol. 7, No. 1/2015

DYNAMICS AND CONTROL OF AN INTEGRO-DIFFERENTIAL SYSTEM OF GEOGRAPHICAL ECONOMICS*

Sebastian Aniţa† Vincenzo Capasso‡ Herb Kunze§
Davide La Torre¶

Dedicated to the memory of Prof. Dr. Viorel Arnăutu

Abstract

In this paper we consider the impact of induced environmental pollution on the qualitative behavior and control of a system of geographical economics. Our underlying mathematical model extends other results in the literature along different directions. A general class of production functions is considered, including, in addition to the classical Cobb-Douglas production function, convex-concave production functions. The dynamics of the pollution is modelled via a diffusion equation coupled, via an integral source, with the geographically distributed production. Reciprocally, we suppose that the (negative) influence of pollution may be modeled as a negative feedback acting on the production function, and therefore on capital accumulation. We analyze the qualitative behavior of the coupled system, and

*Accepted for publication on October 4-th, 2014

†sanita@uaic.ro, Faculty of Mathematics,"Al.I. Cuza" University of Iaşi and "Octav Mayer" Institute of Mathematics, Iaşi, Romania

‡vincenzo.capasso@unimi.it University of Milan, Italy, and University "Carlos III", Madrid, Spain

§hkunze@uoguelph.ca University of Guelph, Department of Mathematics and Statistics, Guelph, Ontario, Canada

¶davide.latorre@unimi.it Khalifa University, Department of Applied Mathematics and Sciences, Abu Dhabi, UAE, and University of Milan, Department of Economics, Management and Quantitative Methods, Milan, Italy

then propose an optimal control problem for the above model. In order to solve the system of partial differential equations which describes the optimality conditions, we implement a Forward-Backward Sweep algorithm. Numerical simulations are reported which illustrate the behavior of the system and its optimal control.

MSC: 35K57; 35Q91; 93D15; 49K20; 91B62; 91B76

keywords: Geographical economics; Environmental pollution; Reaction-diffusion systems; Integral nonlocal term; Qualitative analysis; Optimal control; Non-concave production function

1 Introduction

Different from standard macroeconomic models and environmental economics, recent literature tends to develop a global theory combining these two branches of literature (see [12]). In this paper, following this new trend, we analyze the negative impact of induced pollution on the qualitative behavior, and on the control of a mathematical model of geographical economics. We suppose that the (negative) influence of pollution may be modeled as a negative feedback to the production function and therefore on capital accumulation.

The first studies in geographical economics date back to Beckman [9] and Puu [25], who consider regional problems based simply on flow equations. These works led to the development of a notion of geographical economics that uses general equilibrium models to analyze the peculiarities of local and global markets, as well as the mobility of production factors (see [20], [22], [23]). More recently, this geographical approach has been introduced in economic growth models to study the connections between accumulation and diffusion of capital on economic dynamics (see [10], [11], [14], [16]). The Solow model [30] with a continuous spatial dimension has been extensively studied. Camacho and Zou [14] analyze problems of convergence across regions when capital is mobile, while Brito [11] considers the case in which both capital and labor are mobile. Capasso et al. [16] introduce technology diffusion in the same model, under the additional feature of a convex-concave production function. The Ramsey model [26] has been extended to a spatial dimension by Brito [11] and Boucekkine et al. [10], respectively in average and total utilitarianism versions. Other contributions which explore the spatial dimension in environmental and resource economics can be found in [5], [13], [31].

Before moving to the case analyzed in this paper, we wish to point out a key feature of our models (see also [16]) which concerns the extension of the production function to a larger class of functions, including both the classical Cobb-Douglas case, and convex-concave production functions. The neoclassical theory, developed firstly by Solow (1956), is founded, among others, on two main assumptions driving its main results, namely the fact that the production function exhibits decreasing marginal returns, and infinite marginal productivity for very small input levels (Inada conditions). This ensured that a unique non-trivial equilibrium exists, so that every economic system would converge in the long-run to such a capital level. However, such a model provide a good description of systems with an high level of economic development, and are not applicable to less developed countries (see [27], [28]). In fact, the presence of fixed costs is an importance hindrance to the development of poor countries and only when the production level can get sufficiently high to compensate for such costs, returns will become decreasing. In order to build a theory able to describe the evolution of both advanced and less favored countries, we have relaxed these assumptions while keeping the general framework unaltered; in this way we have shown that it is possible to predict the so called *poverty traps* [16]. In [17] related inverse problems have been faced.

A short announcement of the main results of our research has been presented in a letter [3]; here we offer all relevant mathematical analysis supporting the anticipated results, together with the outcomes of related numerical simulations.

In Section 2 we present the underlying mathematical model describing the strong coupling of the evolution equations for the production and the induced environmental pollution.

In Section 3 we analyze the qualitative behaviour of the system, for large times and for some relevant cases.

In Section 4, we perform a numerical simulation of the steady-state model using reasonable parameter values. The results illustrate that both k and p approach nontrivial and spatially heterogenous equilibria.

In Section 5, we formulate an optimal control problem and solve a particular case using the Forward-Backward Sweep method. We assume that there is a representative agent who wishes to maximize his inter-temporal utility subject to the dynamic constraints (2)-(4). If we denote by $c(x,t)k(x,t)$ the pointwise instantaneous "harvesting effort", the control

problem reads as

$$\max_{c\in U} \int_0^T \int_\Omega e^{-\delta t} c(x,t)k(x,t)dx\ dt, \tag{1}$$

subject to the relevant dynamics; δ is a nonnegative real number. The general cost function using the CIES utility function would have been $\frac{[c(x,t)k(x,t)]^{1-\beta}-1}{1-\beta}$, where $\beta \in [0,1)$ is a positive parameter; we may anticipate that this general case can be treated in the same fashion as we have treated here the case when $\beta = 0$, as in (1).

2 The underlying dynamical model

Let $k(x,t)$ and $p(x,t)$ respectively denote the capital stock held by and the pollution stock faced by a representative household located at x at date t, in a habitat $\overline{\Omega}$ (where $\Omega \subset \mathrm{R}^{\mathbf{2}}$ is taken as a nonempty and bounded domain with a smooth boundary), and $t \geq 0$. We also assume that the initial capital and pollution distributions, $k(x,0) = k_0(x)$ and $p(x,0) = p_0(x)$, are known and satisfy

$$k_0,\ p_0 \in L^\infty(\Omega), \quad k_0(x) \geq k_{00} > 0, p_0(x) \geq 0 \quad \text{a.e. } x \in \Omega, \tag{H}$$

and there is no capital or pollution flow through the boundary of Ω, namely that the normal derivatives $\frac{\partial k}{\partial \nu}(x,t) = \frac{\partial p}{\partial \nu}(x,t) = 0$ at $x \in \partial\Omega$ and $t \geq 0$. We assume a continuous space structure of both physical capital and pollution, so that the model we are interested in is the following:

$$\begin{cases} \dfrac{\partial k}{\partial t}(x,t) = d_1 \Delta k(x,t) + \dfrac{sf(k(x,t))}{g(p)} - \delta_1 k(x,t) - c(x,t)k(x,t) \\ \dfrac{\partial p}{\partial t}(x,t) = d_2 \Delta p(x,t) + \theta \displaystyle\int_\Omega f(k(x',t))\varphi(x',x)dx' - \delta_2 p(x,t) \end{cases} \tag{2}$$

for $(x,t) \in Q_{0,\infty}$, where

$g : [0,+\infty) \to (0,+\infty)$ is continuously differentiable and increasing,

$g(0) = 1$ and $\lim_{r\to+\infty} g(r) = +\infty$,

subject to homogeneous Neumann boundary conditions

$$\frac{\partial k}{\partial \nu}(x,t) = \frac{\partial p}{\partial \nu}(x,t) = 0, \qquad (x,t) \in \Sigma_{0,\infty}, \tag{3}$$

and initial conditions

$$k(x,0) = k_0(x),\ p(x,0) = p_0(x), \qquad x \in \Omega. \tag{4}$$

The control variable $c(x,t)$ describes the level of consumption at the location x, at the time t ($c \in L^\infty(\Omega \times (0,+\infty))$, $0 \leq c(x,t) \leq L$ a.e.) and $d_1, d_2, s, \theta, \delta_1, \delta_2, L$ are positive parameters. Here $Q_{a,b} = \Omega \times (a,b)$ and $\Sigma_{a,b} = \partial\Omega \times (a,b)$.

In the above model (2) the symbol f denotes a production function; we assume it is of the following form

$$f(r) = \frac{\alpha_1 r^\gamma}{1 + \alpha_2 r^\gamma}, \tag{5}$$

where $\alpha_1 \in (0,+\infty)$, $\alpha_2 \in [0,+\infty)$, $\gamma \in (0,+\infty)$. The choice of $g(p) = 1+p^2$ appears suddenly in the literature. We shall use this assumption in our present paper as well.

For basic results concerning the solutions to reaction-diffusion systems without integral terms we refer to [29]. We wish to remark that we deal here with a reaction-diffusion system including an integral (nonlocal) feedback in the evolution equation of the pollution concentration.

Let us notice that for $\alpha_2 = 0$ and $\gamma \in (0,1]$, we get the well known Cobb-Douglas production function. On the other hand, for $\alpha_2 > 0$ and $\gamma > 1$, we get an S-shaped production function. Paper [28] is the first contribution in the economic literature dealing with non-concave or convex/concave production functions. From an economic perspective this kind of assumption is justified by empirical evidences from less developed countries. Finally, the kernel $\varphi(x',x)$ describes the way in which pollution spreads over space; it satisfies the following hypotheses: $\varphi \in L^\infty(\Omega \times \Omega)$, and $\varphi(x',x) \geq 0$, for a.e. $(x',x) \in \Omega \times \Omega$. For $\gamma \in [1,+\infty)$, via Banach's fixed point theorem and using the fact that f is continuously differentiable, it is possible to prove that there exists a unique and nonnegative solution to (2)-(4) on the whole positive time semi-axis. Whenever $\gamma \in (0,1)$, f is no longer differentiable at 0; however, since by (H) we have that $k_0(x) \geq k_{00} > 0$ a.e. $x \in \Omega$, comparison results for parabolic equations and the fixed point theorem imply, in this case too, the existence and uniqueness of a nonnegative solution to (2)-(4), on the whole positive time semi-axis.

3 Large-time behavior of the underlying dynamical system

3.1 The case $p \equiv 0$ and time-independent c

In this case system (2)-(4) reduces to

$$\begin{cases} \dfrac{\partial k}{\partial t} = d_1 \Delta k(x,t) + sf(k(x,t)) - \delta_1 k(x,t) - c(x)k(x,t), & (x,t) \in Q_{0,\infty} \\ \dfrac{\partial k}{\partial \nu}(x,t) = 0, & (x,t) \in \Sigma_{0,\infty} \\ k(x,0) = k_0(x), & x \in \Omega. \end{cases} \tag{6}$$

In the following we discuss the large time behavior of (6) under different hypotheses on the parameter values.

I) In the production function (5) we first assume that $\alpha_2 = 0$ and $\gamma \in (0,1)$. In this case we are assuming that the production function f takes a Cobb-Douglas form. Then for any space independent initial datum k_{01}, with $k_{01} > 0$ sufficiently small, we get

$$sf(k_{01}) - \delta_1 k_{01} - c(x)k_{01} > 0, \quad \text{a.e. } x \in \Omega;$$

i.e. $k_{01}(\cdot)$ is a (strict) lower solution to

$$\begin{cases} d_1 \Delta \widetilde{k}(x) + sf(\widetilde{k}(x)) - \delta_1 \widetilde{k}(x) - c(x)\widetilde{k}(x) = 0, & x \in \Omega \\ \dfrac{\partial \widetilde{k}}{\partial \nu}(x) = 0, & x \in \partial\Omega. \end{cases} \tag{7}$$

Hence, by using comparison results for parabolic equations (see e.g. [19]) we obtain that the solution k_1 to (6), subject to the initial datum k_{01}, is monotonically increasing in $t \in [0, +\infty)$, for almost any $x \in \Omega$.

On the other hand any space independent initial datum k_{02} with $k_{02} > 0$ sufficiently large, is a (strict) upper solution of (7), i.e.

$$sf(k_{02}) - \delta_1 k_{02} - c(x)k_{02} < 0, \quad \text{a.e. } x \in \Omega.$$

By the same arguments as above, we obtain that the solution k_2 to (6), subject to the initial datum k_{02}, is monotonically decreasing in $t \in [0, +\infty)$, for almost any $x \in \Omega$.

The monotonicity of k_1 and k_2 implies that $k_1(\cdot,t) \to \tilde{k}_1$, $k_2(\cdot,t) \to \tilde{k}_2$ in $L^q(\Omega)$, as $t \to +\infty$, for any $q \in [1,+\infty)$, where $\tilde{k}_1$ and $\tilde{k}_2$ are nonnegative solutions to (7). In addition this yields $0 < \tilde{k}_1(x) \leq \tilde{k}_2(x)$ a.e. $x \in \Omega$.

Actually, using a standard argument for parabolic equations (see [29]), let us prove that $\tilde{k}_1(x) = \tilde{k}_2(x)$, a.e. $x \in \Omega$. Since $\tilde{k}_1$ satisfies

$$\begin{cases} d_1 \Delta \tilde{k}_1(x) + s\alpha_1 \tilde{k}_1(x)^\gamma - \delta_1 \tilde{k}_1(x) - c(x)\tilde{k}_1(x) = 0, & x \in \Omega \\ \dfrac{\partial \tilde{k}_1}{\partial \nu}(x) = 0, & x \in \partial\Omega, \end{cases}$$

multiplying by $\tilde{k}_2$ and integrating on Ω gives that

$$-d_1 \int_\Omega \nabla \tilde{k}_1 \nabla \tilde{k}_2 dx + s\alpha_1 \int_\Omega \tilde{k}_1^\gamma \tilde{k}_2 dx = \int_\Omega (\delta_1 + c)\tilde{k}_1 \tilde{k}_2 dx.$$

In the same manner we get that

$$-d_1 \int_\Omega \nabla \tilde{k}_1 \nabla \tilde{k}_2 dx + s\alpha_1 \int_\Omega \tilde{k}_1 \tilde{k}_2^\gamma dx = \int_\Omega (\delta_1 + c)\tilde{k}_1 \tilde{k}_2 dx.$$

We infer that

$$\int_\Omega \tilde{k}_1(x)^\gamma \tilde{k}_2(x)^\gamma (\tilde{k}_2(x)^{1-\gamma} - \tilde{k}_1(x)^{1-\gamma}) dx = 0$$

and taking into account that $\tilde{k}_1$ and $\tilde{k}_2$ are positive and $\tilde{k}_1(x) \leq \tilde{k}_2(x)$ a.e. $x \in \Omega$, we conclude that $\tilde{k}_1 \equiv \tilde{k}_2$. Let us denote by $\tilde{k}(x)$, a.e. $x \in \Omega$, the common function.

We may now notice that, for any k_0 satisfying (H) we can choose the space independent $k_{01} > 0$ sufficiently small and $k_{02} > 0$ sufficiently large and such that $k_{01} \leq k_0(x) \leq k_{02}$ for a.e. $x \in \Omega$. Again the comparison results in [19] imply that any solution k to (6) subject to the initial datum k_0 satisfies $\lim_{t\to+\infty} k(\cdot,t) = \tilde{k}$ in $L^2(\Omega)$. Regularity results for the solutions of parabolic equations imply that $\lim_{t\to+\infty} k(\cdot,t) = \tilde{k}$ in $L^\infty(\Omega)$ as well [4].

II) In the production function, see (5), we assume that $\alpha_2 = 0$ and $\gamma \in (1,+\infty)$. For any space independent $k_{01} > 0$, with k_{01} sufficiently small, we get that for any $t \in (0,+\infty)$ sufficiently small :

$$\begin{aligned} & sf(k_1(x,t)) - \delta_1 k_1(x,t) - c(x)k_1(x,t) \\ = \;& s\alpha_1 k_1(x,t)^\gamma - \delta_1 k_1(x,t) - c(x)k_1(x,t) \;\leq\; -\frac{\delta_1}{2} k_1(x,t) \end{aligned}$$

a.e. $x \in \Omega$, where k_1 is the solution to (6) corresponding to $k_0 := k_{01}$. The comparison result for parabolic equations implies that the mapping $t \mapsto k_1(x,t)$ is decreasing on $[0,+\infty)$ for almost any $x \in \Omega$ and consequently

$$\begin{aligned} & sf(k_1(x,t)) - \delta_1 k_1(x,t) - c(x)k_1(x,t) \\ = \quad & s\alpha_1 k_1(x,t)^\gamma - \delta_1 k_1(x,t) - c(x)k_1(x,t) \ \leq \ -\frac{\delta_1}{2}k_1(x,t) \end{aligned}$$

for any $t \in [0,+\infty)$, a.e. $x \in \Omega$. We then deduce that

$$0 < k_1(x,t) \leq k_{11}(x,t)$$

a.e. $x \in \Omega$, for any $t \in [0,+\infty)$, where k_{11} is the solution to

$$\begin{cases} \dfrac{\partial k}{\partial t}(x,t) = d_1 \Delta k(x,t) - \dfrac{\delta_1}{2}k(x,t), & (x,t) \in Q_{0,\infty} \\ \dfrac{\partial k}{\partial \nu}(x,t) = 0, & (x,t) \in \Sigma_{0,\infty} \\ k(x,0) = k_{01}, & x \in \Omega. \end{cases} \tag{8}$$

Since the unique solution to (8) is $k_{11}(x,t) = k_{01}\exp\{-\frac{\delta_1 t}{2}\}$, $\forall t \geq 0$, a.e. $x \in \Omega$, we conclude that

$$k_1(\cdot,t) \to 0 \quad \text{in } L^\infty(\Omega),$$

as $t \to +\infty$, exponentially.

On the other hand for any space independent and sufficiently large $k_{02} > 0$, we get in the same manner as in (I1) that for any $t \in [0,+\infty)$:

$$\begin{aligned} & sf(k_2(x,t)) - \delta_1 k_2(x,t) - c(x)k_2(x,t) \\ & = k_2(x,t)[s\alpha_1 k_2(x,t)^{\gamma-1} - \delta_1 - c(x)] \geq \zeta k_2(x,t) \end{aligned}$$

a.e. $x \in \Omega$, where k_2 is the solution to (6) corresponding to $k_0 := k_{02}$ and ζ is a positive constant, and that the mapping $t \mapsto k_2(x,t)$ is increasing on $[0,+\infty)$ for almost any $x \in \Omega$.

Using again the comparison result for parabolic equations we get that

$$k_2(x,t) \geq k_{02}\exp\{\zeta t\},$$

a.e. $x \in \Omega$, for any $t \in [0,+\infty)$, and consequently

$$Ess\ inf_\Omega\ k_2(\cdot,t) \to +\infty,$$

as $t \to +\infty$, exponentially.

III) Let us now assume that $\alpha_2 = 0$ and $\gamma = 1$. Then system (6) becomes

$$\begin{cases} \dfrac{\partial k}{\partial t}(x,t) = d_1 \Delta k(x,t) + (s\alpha_1 - \delta_1 - c(x))k(x,t), & (x,t) \in Q_{0,\infty} \\ \dfrac{\partial k}{\partial \nu}(x,t) = 0, & (x,t) \in \Sigma_{0,\infty} \\ k(x,0) = k_0(x), & x \in \Omega. \end{cases}$$

We have now the following cases happening.

1. For any c satisfying $s\alpha_1 - \delta_1 - c(x) \leq -\zeta$ a.e. $(x,t) \in Q_{0,\infty}$, where ζ is a positive constant, then the comparison result for parabolic equations implies that
$$k(x,t) \leq \|k_0\|_{L^\infty(\Omega)} \exp\{-\zeta t\},$$
a.e. $x \in \Omega$, $\forall t \geq 0$, and so
$$k(\cdot,t) \to 0 \quad \text{in } L^\infty(\Omega),$$
as $t \to +\infty$, exponentially.
2. For any c satisfying $s\alpha_1 - \delta_1 - c(x) \geq \zeta$ a.e. $(x,t) \in Q_{0,\infty}$, where ζ is a positive constant, then the comparison result for parabolic equations implies that
$$k(x,t) \geq k_{00} \exp\{\zeta t\},$$
a.e. $x \in \Omega$, $\forall t \geq 0$, and so
$$Ess\ inf_\Omega\ k(\cdot,t) \to +\infty \quad \text{in } L^\infty(\Omega),$$
as $t \to +\infty$, exponentially.
3. For any c a constant satisfying $s\alpha_1 - \delta_1 - c = 0$, then
$$k(\cdot,t) \to \int_\Omega k_0(x)dx \quad \text{in } L^\infty(\Omega),$$
as $t \to +\infty$.

IV) We consider now the case $\alpha_2 > 0$, with $\gamma \in (1, +\infty)$, so that the production function f, as defined by (5), is $S-$shaped. Then there exists a positive constant η such that $\eta = \sup_{r>0} \dfrac{sf(r)}{r}$. We also have that $f'(r) > 0$, $\forall r > 0$. In addition, if $r \geq 0$ is small then $f(r) \approx \alpha_1 r^\gamma$. If $r > 0$ is large, then $f(r) \approx \frac{\alpha_1}{\alpha_2}$. By the same comparison techniques used above, it is possible to prove the following

1. If we also have that $\eta - \delta_1 - c(x) \leq -c_1 < 0$ a.e. $x \in \Omega$ (where c_1 is a positive constant), then $k(\cdot, t) \to 0$ in $L^\infty(\Omega)$, as $t \to +\infty$.
2. If we assume that $\eta - \delta_1 - c(x) \geq c_1 > 0$ a.e. $x \in \Omega$ (where c_1 is a positive constant), then for any initial datum k_0 with a sufficiently small norm (in $L^\infty(\Omega)$), one gets $k(\cdot, t) \to 0$ in $L^\infty(\Omega)$, as $t \to +\infty$. On the other hand for any initial datum k_0 such that $\dfrac{sf(k_{00})}{k_{00}} \geq \delta_1 + \|c\|_{L^\infty(\Omega)}$ a.e. $x \in \Omega$, we may conclude that $k(\cdot, t) \to \tilde{k}$ in $L^\infty(\Omega)$, as $t \to +\infty$, and $\tilde{k}$ is a nontrivial nonnegative solution to

$$\begin{cases} d_1 \Delta \tilde{k} + s\frac{\alpha_1 \tilde{k}(x)^\gamma}{1+\alpha_2 \tilde{k}(x)^\gamma} - \delta_1 \tilde{k} - c(x)\tilde{k} = 0, & x \in \Omega \\ \dfrac{\partial \tilde{k}}{\partial \nu}(x) = 0, & x \in \partial\Omega. \end{cases} \tag{9}$$

V) Assume now that $\alpha_2 > 0$ and $\gamma = 1$. Then the derivative of

$$G(x, r) := s\frac{\alpha_1 r}{1 + \alpha_2 r} - \delta_1 r - c(x)r$$

with respect to r is

$$\frac{\partial G}{\partial r}(x, r) = s\frac{\alpha_1}{(1 + \alpha_2 r)^2} - \delta_1 - c(x)$$

which is a decreasing function of r.

1. If $\dfrac{\partial G}{\partial r}(x, 0) = s\alpha_1 - \delta_1 - c(x) \leq -c_0 < 0$ a.e. $x \in \Omega$ (c_0 is a positive constant), then the solution k to (2) satisfies

$$k(\cdot, t) \to 0 \qquad \text{in } L^\infty(\Omega)$$

as $t \to +\infty$.

2. If $\dfrac{\partial G}{\partial r}(x, 0) \geq c_0 > 0$ a.e. $x \in \Omega$, then it follows as in the case (I1) that for any space independent and sufficiently small $k_{01} > 0$ we get that

$$k_1(\cdot, t) \to \tilde{k}_1 \quad \text{in } L^\infty(\Omega),$$

as $t \to +\infty$, where k_1 is the solution to (6) corresponding to $k_0 := k_{01}$, and $\tilde{k}_1$ is a positive solution to (7). In the

same manner as in case (I) it also follows that $\tilde{k}_1 = \tilde{k}_2 = \tilde{k}$, which is the unique nontrivial nonnegative solution to (7), where k_2 and $\tilde{k}_2$ are constructed as in (I). Using again the comparison result for parabolic equations we get that

$$k(\cdot, t) \to \tilde{k} \quad \text{in } L^\infty(\Omega),$$

as $t \to +\infty$, where k is the solution to (6).

VI) If $\alpha_2 > 0$ and $\gamma \in (0, 1)$, then for any space independent and sufficiently small $k_{01} > 0$ we get that the mapping $t \mapsto k_1(x, t)$ is increasing on $[0, +\infty)$, for almost any $x \in \Omega$ and that

$$k_1(\cdot, t) \to \tilde{k}_1 \quad \text{in } L^\infty(\Omega),$$

as $t \to +\infty$, where k_1 is the solution to (6) corresponding to $k_0 := k_{01}$ and $\tilde{k}_1$ is a solution to (7) satisfying in addition

$$0 < \tilde{k}_1(x) \leq \tilde{k}_2(x)$$

a.e. $x \in \Omega$. Here k_2 and $\tilde{k}_2$ are constructed as in (I). As in the first case (I) we get that

$$\int_\Omega (\tilde{k}_2(x) f(\tilde{k}_1(x)) - \tilde{k}_1(x) f(\tilde{k}_2(x)))dx = 0$$

and since the function integrated here is nonnegative, we may conclude that

$$\tilde{k}_2(x) f(\tilde{k}_1(x)) - \tilde{k}_1(x) f(\tilde{k}_2(x)) = 0$$

a.e. $x \in \Omega$ and consequently that $\tilde{k}_1 = \tilde{k}_2 = \tilde{k}$ and this is the unique nontrivial nonnegative solution to (7).

Repeating the argument in case (I) we may finally infer that for any k_0 satisfying (H) we get

$$k(\cdot, t) \to \tilde{k} \quad \text{in } L^\infty(\Omega),$$

as $t \to +\infty$.

The above discussion shows that, under the hypotheses $\alpha_2 > 0$ and $\gamma > 1$, which correspond to the case of an S-shaped production function, a saddle behavior emerges; i.e. for sufficiently small initial datum k_0, the production

$k(x,t)$ diminishes to 0, as $t \to +\infty$, while when k_0 is sufficiently large the production tends to a certain nontrivial steady state. As a conclusion, let us notice that these results could also be obtained by assuming a general function $f \in C^1([0,+\infty))$ such that $f'(0) = 0$, $f'(r) > 0$ for any $r > 0$ and $\lim_{r\to+\infty} f(r) = \tau \in (0,+\infty)$ (see e.g. [15]). To investigate systems with nonlinear diffusion we have to combine the techniques in this paper with those in [18].

3.2 The general case with pollution diffusion

We are dealing here with the case when $\alpha_2 > 0$ and $\gamma > 1$. Assume that $0 \le c(x,t) \le L$ a.e. $(x,t) \in Q_{0,\infty}$. Let (k,p) be the solution to (2)-(4). Comparison results for parabolic equations imply that $k(x,t) \le \overline{k}_2(x,t)$ a.e. $(x,t) \in Q_{0,\infty}$, where $\overline{k}_2$ is the solution to (6) corresponding to $c \equiv 0$ and $p \equiv 0$. By using comparison results for parabolic equations including integral terms (see [2]), we get that $0 \le p(x,t) \le \overline{p}_2(x,t)$ a.e. $(x,t) \in Q_{0,\infty}$, where $\overline{p}_2$ is the solution to

$$\frac{\partial \overline{p}_2}{\partial t}(x,t) = d_2\Delta\overline{p}_2(x,t) + \theta\int_\Omega f(\overline{k}_2(x',t))\varphi(x',x)dx' - \delta_2\overline{p}_2(x,t),\ (x,t) \in Q_{0,\infty},$$

subject to boundary and initial conditions as in (3) and (4).

If k_{00} is sufficiently large, then $\overline{k}_2(t) \to \tilde{k}_2$ in $L^\infty(\Omega)$, as $t \to +\infty$, where $\tilde{k}_2$ is the maximal nonnegative solution to (5) corresponding to $c \equiv 0$. This implies that $\overline{p}_2(\cdot,t) \to \tilde{p}_2$ in $L^\infty(\Omega)$, as $t \to +\infty$, where $\tilde{p}_2$ is the solution to

$$\begin{cases} d_2\Delta\tilde{p}_2(x) + \theta\displaystyle\int_\Omega f(\tilde{k}_2(x'))\varphi(x',x)dx' - \delta_2\tilde{p}_2(x) = 0, & x \in \Omega \\ \dfrac{\partial\tilde{p}_2}{\partial\nu}(x) = 0, & x \in \partial\Omega. \end{cases}$$

Now for any $\varepsilon > 0$, there exists $t(\varepsilon) > 0$ such that $k(x,t) \ge k^*_\varepsilon(x,t)$ a.e. $x \in \Omega$, for all $t \ge t(\varepsilon)$, where k^*_ε is the solution to

$$\begin{cases} \dfrac{\partial k^*_\varepsilon}{\partial t} = d_1\Delta k^*_\varepsilon + \dfrac{sf(k^*_\varepsilon(x,t))}{g(\tilde{p}_2(x)+\varepsilon)} - \delta_1 k^*_\varepsilon(x,t) - Lk(x,t), & (x,t) \in Q_{t(\varepsilon),\infty} \\ \dfrac{\partial k^*_\varepsilon}{\partial\nu}(x,t) = 0, & (x,t) \in \Sigma_{t(\varepsilon),\infty}, \end{cases}$$

satisfying $k^*_\varepsilon(x,t(\varepsilon)) = k(x,t(\varepsilon))$ a.e. $x \in \Omega$. In conclusion, if $\frac{\eta}{g(\tilde{p}_2(x))} - \delta_1 - L \ge \mu > 0$ a.e. $x \in \Omega$, then for any k_{00} sufficiently large we get the existence a sustainable economy, characterized by the persistence of k and the boundedness of the level of pollution. Moreover, for k_0 sufficiently small, the production $k(\cdot,t)$ tends to 0 in $L^\infty(\Omega)$, as $t \to +\infty$, which corresponds to a collapsing economy.

4 Numerical simulations

In the following simulations we use the parameter values

$$\begin{cases} \delta_1 = 0.05,\ \delta_2 = 0.01,\ s = 0.25,\ d_1 = 0.01,\ d_2 = 0.01,\ \theta = 0.1,\ c \equiv 0 \\ \varphi(x', x) = \frac{1}{\sqrt{\pi\varepsilon}} e^{-\frac{|x-x'|^2}{\varepsilon}} \psi(x),\ \varepsilon = 0.001,\ \psi(x) = x^2 \\ k(x, 0) = e^{-x^2} \text{ and } p(x, 0) = e^x,\ \Omega = [a, b] = [-1, 1],\ T = 600 \\ \alpha_1 = 100,\ \alpha_2 = 100,\ \gamma = 4 \\ g(p) = 1 + p^2. \end{cases} \tag{10}$$

The above choices of parameter values are explained as follows:

- $\delta_1 = 0.05$ can be found in [6] and they describe the physical capital share and the depreciation rate of physical capital, respectively.
- $\delta_2 = 0.01$ represents the environmental ability to absorb pollution. The growth of CO2 emissions tripled between 2000 and 2004, growing by more than 3 percent per year according to a new study published in Proceedings of the National Academy of Sciences USA. Since the air quality is decreasing, it is ural to suppose that the environmental ability to absorb pollution is less than 3 per cent.
- d_1 and d_2 determine the diffusivity. We set them both equal to 0.01.
- s is an efficiency parameter that we assume to be greater than 0.2 (see [21]).
- θ is a normalization factor.
- In $\varphi(x', x)$, the expression $\frac{1}{\sqrt{\pi\varepsilon}} e^{-\frac{|x-x'|^2}{\varepsilon}}$ is a classical Gaussian kernel, and the function $\psi(x)$ allows from some place-dependent behaviour in the kernel.
- $\alpha_1 = \alpha_2 = 100$ and $\gamma = 4$ are set to values so that f is S-shaped.
- $T = 600$ is the length of the time interval.

The solution surfaces are plotted in Figure 1.

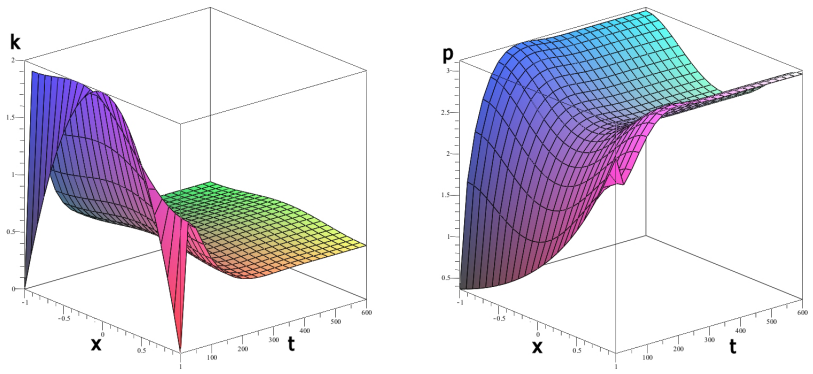

Figure 1: Solution surfaces for capital and pollution in the case of an S-shaped production function, kernel, and parameter values as in (10).

5 An optimal harvesting problem

Here we will consider only the case $\beta = 0$, $\alpha_2 > 0$, $\gamma > 1$. The results in the previous section imply that for any $c \in L^\infty(Q_{0,\infty})$, $0 \le c(x,t) \le L$ a.e. in Q, then (k^c, p^c), the solution to (2)-(4) corresponding to c satisfies $0 \le k^c(x,t) \le \overline{k}_2(x,t)$ a.e. $(x,t) \in Q$. Since $\overline{k}_2 \in L^\infty(Q_{0,\infty})$, there exists $M \ge 0$ such that $0 \le \overline{k}_2(x,t) \le M$ a.e. $(x,t) \in Q_{0,\infty}$. We may conclude that

$$0 \le \int_0^\infty \int_\Omega e^{-\delta t} c(x,t) k^c(x,t) dx \; dt \le LMmeas(\Omega)\frac{1}{\delta}$$

and

$$0 \le \int_T^\infty \int_\Omega e^{-\delta t} c(x,t) k^c(x,t) dx \; dt \le LMmeas(\Omega)\frac{e^{-\delta T}}{\delta}.$$

This means that instead of investigating the control problem formulated in Section 2 we could treat the following approximating optimal control problem with a finite horizon time (it is an optimal harvesting problem):

$$(OH) \qquad \max_{c \in U} \int_0^T \int_\Omega e^{-\delta t} c(x,t) k^c(x,t) dx \; dt,$$

where $T > 0$ is fixed (and large), and $U = \{v \in L^\infty(Q_{0,T});\ 0 \le v(x,t) \le L \text{ a.e. in } Q_{0,T}\}$ is the set of controls, and (k^c, p^c) is the solution to (2)-(4) corresponding to $g(p) = 1 + p^2$, and $Q_{0,T}$ and $\Sigma_{0,T}$ (instead of $Q_{0,\infty}$ and $\Sigma_{0,\infty}$, respectively).

Since this is a standard optimal control problem, the existence of at least one optimal control c^* can be proven following [1, 7, 8]. In addition, the following result holds.

Theorem 1 *If (k^*, p^*) is the optimal state corresponding to c^*, and if (q_1, q_2) is the solution to the following problem*

$$\begin{cases} \dfrac{\partial q_1}{\partial t}(x,t) = -d_1\Delta q_1(x,t) - \left(s\dfrac{f'(k^*(x,t))}{1+p^*(x,t)^2} - \delta_1\right) q_1(x,t) \\ \qquad -\theta f'(k^*(x,t)) \int_\Omega q_2(x',t)\varphi(x,x')dx' + c^*(x,t)(e^{-\delta t} + q_1(x,t)) \\ \dfrac{\partial q_2}{\partial t}(x,t) = -d_2\Delta q_2(x,t) + \dfrac{2sf(k^*(x,t))p^*(x,t)}{(1+p^*(x,t)^2)^2} q_1(x,t) + \delta_2 q_2(x,t), \end{cases} \tag{11}$$

for $(x,t) \in Q_{0,T}$, subject to homogeneous Neumann boundary conditions and final conditions

$$q_1(x,T) = 0,\ q_2(x,T) = 0, \qquad x \in \Omega, \tag{12}$$

then

$$c^*(x,t) = \begin{cases} 0, & if\ e^{-\delta t} + q_1(x,t) < 0 \\ L, & if\ e^{-\delta t} + q_1(x,t) > 0. \end{cases} \tag{13}$$

Equations (11)-(13) provide the necessary optimality conditions for (OH). It is obvious that taking into account (13) we may rewrite (11) as

$$\begin{cases} \dfrac{\partial q_1}{\partial t}(x,t) = -d_1\Delta q_1(x,t) - \left(s\dfrac{f'(k^*(x,t))}{1+p^*(x,t)^2} - \delta_1\right) q_1(x,t) \\ \qquad -\theta f'(k^*(x,t)) \int_\Omega q_2(x',t)\varphi(x,x')dx' + L(e^{-\delta t} + q_1(x,t))^+, \\ \dfrac{\partial q_2}{\partial t}(x,t) = -d_2\Delta q_2(x,t) + \dfrac{2sf(k^*(x,t))p^*(x,t)}{(1+p^*(x,t)^2)^2} q_1(x,t) + \delta_2 q_2(x,t), \end{cases} \tag{14}$$

for $(x,t) \in Q_{0,T}$, subject to homogeneous Neumann boundary conditions, and final conditions (12). Using the theorem given before, we can derive a gradient type algorithm (see [1]) to approximate the optimal control c^*.

We must solve Equations (2) and (14) subject to homogeneous Neumann boundary conditions, the initial conditions at $t = 0$ for $k(x,t)$ and $p(x,t)$, and the final conditions at $t = T$ for $q_1(x,t)$ and $q_2(x,t)$. One solution approach is to reverse time in Equations (14) via the change of variable $\tau = T - t$, turning the problem for q_1 and q_2 into forward problem with zero initial conditions. Starting with the solutions $k^0(x,t)$ and $p^0(x,t)$ corresponding to $c^0(x,t) = 0$, we use an iterative procedure.

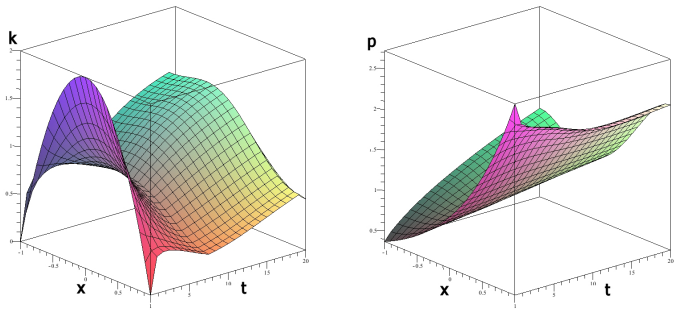

Figure 2: Approximations of the optimal states k^* and p^*

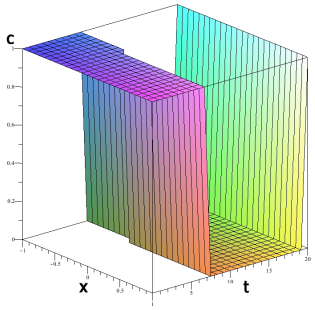

Figure 3: Approximation of the optimal control c^*.

The iterative algorithm is very intuitive, and efficient. It is generally referred to as the Forward-Backward Sweep method [24]. A numerical simulation of the above procedure is provided in the following example.

Example: We solve the optimal control problem using the same parameters as in Section 4, but for $L = 1$, and $T = 20$, and in addition setting $\delta = 0.1$ in the objective function. In Figure 2 we plot the optimal states k^* and p^*. We also include an approximation of the level sets of c^* in Figure 3.

Acknowledgement. The work of S. Aniţa was supported by a grant of the Ministry of National Education, CNCS-UEFISCDI, project number PN-II-ID-PCE-2012-4-0270 (68/2.09.2013): "Optimal Control and Stabilization of Nonlinear Parabolic Systems with State Constraints. Applications in Life Sciences and Economics".

References

[1] S. Aniţa, V. Arnăutu, V. Capasso, *An introduction to optimal control problems in life sciences and economics. From mathematical models to numerical simulation with MATLAB,* Birkhäuser, Berlin, 2010.

[2] S. Aniţa, V. Capasso, A stabilization strategy for a reaction-diffusion system modelling a class of spatially structured epidemic systems (think globally, act locally), *Nonlin. Anal. Real World Appl.* 10:2026-2035, 2009.

[3] S. Aniţa, V. Capasso, H. Kunze, D. La Torre, Optimal control and long-run dynamics for a spatial economic growth model with physical capital accumulation and pollution diffusion, *Appl. Math. Lett.* 26:908-912, 2013.

[4] S. Aniţa, D. Tătaru, Null controllability for the dissipative semilinear heat equation, *Appl. Math. Optimiz.* 46:97-105, 2002.

[5] S. Athanassoglou, A. Xepapadeas, Pollution control with uncertain stock dynamics: When, and how, to be precautious, *J. Environ. Econ. Manag.* 63:304-320, 2012.

[6] R.J. Barro, X.I. Sala-i-Martin, *Economic Growth,* MIT Press, Boston, 2003.

[7] V. Barbu, *Analysis and control of nonlinear infinite dimensional systems,* Academic Press Inc., Boston, 1993.

[8] V. Barbu, *Mathematical methods in optimization of differential systems,* Kluwer Acad. Publ., Dordrecht, 1994.

[9] W. Beckerman, Economic growth and the environment: Whose growth? Whose environment?, *World Dev.* 20:481-496, 1992.

[10] R. Boucekkine, C. Camacho, B. Zou, Bridging the gap between growth theory and the new economic geography: the spatial Ramsey model, *Macroecon. Dyn.* 13:20-45, 2009.

[11] P. Brito, The dynamics of growth and distribution in a spatially heterogeneous world, UECE-ISEG, Technical University of Lisbon, 2004.

[12] W.A. Brock, M.S. Taylor, The green Solow model, *J. Econ. Growth* 15:127-153, 2010.

[13] W. Brock, A. Xepapadeas, Pattern formation, spatial externalities and regulation in coupled economic-ecological systems, *J. Environ. Econ. Manag.* 59:149-164, 2010.

[14] C. Camacho, B. Zou, The spatial Solow model, *Econ. Bull.* 18:1-11, 2004.

[15] V. Capasso, *Mathematical structures of epidemic systems,* Lecture Notes in Biomathematics, Vol. 97. Springer-Verlag, Heidelberg, 1993. Second corrected printing, 2008.

[16] V. Capasso, R. Engbers, D. La Torre, On the spatial Solow model with technological diffusion and nonconcave production function, *Nonlin. Anal. Real World Appl.* 11:3858-3876, 2010.

[17] R. Engbers, M. Burger, V. Capasso, Inverse problems in geographical economics: parameter identification in the spatial Solow model, *Proc. Royal Soc. A*, in press, 2014.

[18] A. Favini, G. Marinoschi, *Degenerate nonlinear diffusion equations,* Lecture Notes in Mathematics 2049, Springer, Berlin, 2012.

[19] A. Friedman, *Partial differential equations of parabolic type,* Prentice-Hall, Englewood Cliffs, N.J., 1964.

[20] M. Fujita, P. Krugman, A. Venables, *The spatial economy. Cities, regions and international trade,* MIT Press, Cambridge, 1999.

[21] R. Klump, P. McAdam, A. Willman, The normalized CES production function: theory and empirics, European Central Bank, Working Paper Series N. 1294, February 2011.

[22] P. Krugman, Increasing returns and economic geography, *J. Polit. Econ.* 99:483-499, 1991.

[23] P. Krugman, *Development, geography and economic theory,* MIT Press, Cambridge (MA), 1995.

[24] S. Lenhart, J.T. Workman, Optimal control applied to biological models, Chapman & Hall/CRC, Boca Raton, 2007.

[25] T. Puu, Outline of a trade cycle model in continuous space and time, *Geogr. Anal.* 14:1-9, 1982.

[26] F.P. Ramsey, A mathematical theory of saving, *Econ. J.* 38:543-559, 1928.

[27] J.D. Sachs, J.W. McArthur, G. Schimdt-Traub, M. Kruk, C. Bahadur, M. Faye, G. McCord, Ending Africa's poverty trap, *Brookings Papers Econ. Activ.* 1:117-240, 2004.

[28] A. K. Skiba, Optimal growth with a convex-concave production function, *Econometrica* 46:527-539, 1978.

[29] J. Smoller, *Shock waves and reaction-diffusion equations,* Springer-Verlag, Berlin, Heidelberg, 1983.

[30] R.M. Solow, A contribution to the theory of economic growth, *Quart. J. Econ.* 70:65-94, 1956.

[31] A. Xepapadeas, The spatial dimension in environmental and resource economics, *Environ. Develop. Econ.* 15:747-758, 2010.

Ann. Acad. Rom. Sci.
Ser. Math. Appl.
ISSN 2066-6594 Vol. 7, No. 1/2015

OPTIMIZING THE POSITION OF THE SUPPORT OF THE CONTROL FOR SOME OPTIMAL HARVESTING PROBLEMS*

Ana-Maria Moşneagu†

Dedicated to the memory of Prof. Dr. Viorel Arnăutu

Abstract

In this paper we investigate the optimal position of the support of the control for some optimal harvesting problems. First we refer to a logistic model with diffusion. We remind the existence result of an optimal control and the necessary optimality conditions for the related optimal harvesting problem. Then we obtain an iterative method to improve the position of the support of the optimal harvesting effort in order to maximize the harvest (for a simplified model without logistic term). Numerical tests illustrating the effectiveness of the theoretical results are given.

MSC: 92D25; 49K20; 65M06; 35Q92

Keywords: Population dynamics; Diffusive models; Optimal control; Numerical methods.

1 Introduction

Since R. A. Fisher introduced in [12] a mathematical model of spatially structured population, a related flourishing literature was developed (e.g.

*Accepted for publication on November 28-th, 2014

†anamaria.mosneagu@uaic.ro, Faculty of Mathematics, "Alexandru Ioan Cuza" University of Iaşi, 700506 Iaşi, Romania

[14], [15], [16],) which combines diffusive motion of individuals with nonlinearities arising from their growth and competition process. For models related to dynamics of population we refer to [3]. In this paper we recall the results obtained in [8] for a logistic model with diffusion. We consider a related optimal harvesting problem. We want to find firstly the magnitude of the control that acts on a certain subdomain and to study the position of the subdomain where the control acts in order to optimize the cost (for basic results and methods in the optimal shape design theory we refer to [13]).

We consider the following Fisher's model corresponding to a biological population that is free to move in an isolated habitat $\Omega \subset \mathbf{R}^N,\ N \in \{2,3\}$:

$$\begin{cases} \partial_t y(x,t) - d\Delta y(x,t) = a(x)y(x,t) - k(x)y^2(x,t), & (x,t) \in Q_T, \\ \partial_\nu y(x,t) = 0, & (x,t) \in \Sigma_T, \\ y(x,0) = y_0(x), & x \in \Omega, \end{cases} \tag{1}$$

where Ω is a domain with a sufficiently smooth boundary $\partial\Omega$, $Q_T := \Omega \times (0,T)$, $\Sigma_T := \partial\Omega \times (0,T)$, $T > 0$, $y = y(x,t)$ is the population density at $(x,t) \in \overline{\Omega} \times [0,T]$ and $y_0(x)$ is the initial population density. The logistic term, $k(x)y^2(x,t)$, describes a local intraspecific competition for resources. Here d is the diffusion coefficient and $a(x)$ indicates the natural growth rate of the population. We have prescribed homogeneous Neumann conditions on the boundary $\partial\Omega$, corresponding to the case of isolated populations. This is an extended model of the one in Section 5.2 from [4], because the population coefficients become functions of x. We start with the following hypotheses:

(H1) $a \in L^\infty(\Omega), \quad d \in (0, +\infty)$;

(H2) $y_0 \in L^\infty(\Omega), \quad y_0(x) \geq 0 \quad$ a.e. $x \in \Omega$ with $\|y_0\|_{L^\infty(\Omega)} > 0$;

(H3) $k \in L^\infty(\Omega), \quad k(x) \geq 0 \quad$ a.e. $x \in \Omega$.

The optimal harvesting problem is

$$Maximize \int_0^T \int_\omega u(x,t)y^u(x,t)dx\ dt, \tag{2}$$

subject to $u \in K$, where $K = \{w \in L^\infty(\omega \times (0,T));\ 0 \leq w(x,t) \leq L \text{ a.e.}\}$, $L > 0$. $u(x,t)$ represents the harvesting effort at $(x,t) \in \omega \times [0,T]$, where $\omega \subset \Omega$ is a nonempty domain with sufficiently smooth boundary $\partial\omega$

and such that $\Omega \setminus \overline{\omega}$ is a domain. Here y^u is the solution of the problem

$$\begin{cases} \partial_t y - d\Delta y = a(x)y - k(x)y^2 - \chi_\omega(x)u(x,t)y(x,t), & (x,t) \in Q_T, \\ \partial_\nu y(x,t) = 0, & (x,t) \in \Sigma_T, \\ y(x,0) = y_0(x), & x \in \Omega, \end{cases} \tag{3}$$

(χ_ω is the characteristic function of ω).
We intend to use the necessary optimality conditions to find the position of ω in Ω (in the set of all of its translations) which gives the maximum value for the harvest. So we have two maximizing problems: firstly, for a fixed ω we find the harvesting effort which gives the maximum harvest; secondly, using this optimal effort (control) we investigate the best position of ω in order to maximize the harvest.

In fact, our problem of optimal harvesting takes the following form:

$$Maximize_{\omega \in \mathcal{O}} Maximize_{u \in K} \int_0^T \int_\omega u(x,t)y^u(x,t)dx\ dt, \tag{4}$$

where $\mathcal{O}$ denotes the set of all translations of ω in Ω.

The paper is structured as follows: in the second section we recall the necessary optimality conditions for our boundary value problem with logistic term and we find the derivative of the optimal cost value with respect to translations of ω in Ω for the linear problem (see [8]). In section 3, we use these results to develop a conceptual iterative algorithm suitable for improving the position of the support of the control. In the last section numerical test are included to sustain the theoretical results.

2 An iterative method to improve the position of the support of the harvesting effort. The case $k \equiv 0$ (the model without logistic term)

First, we refer to the model with logistic term. The existence result of an optimal control for the problem (2) follows the lines in [4].

Theorem 1 *Problem (2) admits at least one optimal control.*

Let us denote by $p = p(x,t)$ the adjoint state, i.e. p satisfies

$$\begin{cases} \partial_t p + d\Delta p = -a(x)p + 2k(x)y^{u^*}p + \chi_\omega(x)u^*(1+p), & (x,t) \in Q_T, \\ \partial_\nu p(x,t) = 0, & (x,t) \in \Sigma_T, \\ p(x,T) = 0, & x \in \Omega, \end{cases} \tag{5}$$

where (u^*, y^{u^*}) is an optimal pair for (2). For the construction of the adjoint problems in optimal control theory we refer to [11]. We have

Theorem 2 *If (u^*, y^{u^*}) is an optimal pair for problem (2) and if p is the solution of problem (5), then we have:*

$$u^*(x,t) = \begin{cases} 0, & 1 + p(x,t) < 0 \\ L, & 1 + p(x,t) > 0 \end{cases} \quad a.e.\ (x,t) \in \omega \times (0,T). \tag{6}$$

Proof. The existence and uniqueness of the adjoint state p can be proved via Banach's fixed point theorem.
Let $v \in L^\infty(\omega \times (0,T))$, arbitrary but fixed, such that $u^* + \varepsilon v \in K$ for sufficiently small $\varepsilon > 0$.
From the optimality of u^* we get that

$$\int_0^T \int_\omega u^*(x,t) \frac{y^{u^*+\varepsilon v}(x,t) - y^{u^*}(x,t)}{\varepsilon} dxdt + \int_0^T \int_\omega v(x,t) y^{u^*+\varepsilon v}(x,t) dxdt \leq 0, \tag{7}$$

for sufficiently small $\varepsilon > 0$.
In order to continue the proof of the theorem, we need the following convergence result (see [4]).

Lemma 1 *One has*

$$y^{u^*+\varepsilon v} \to y^{u^*} \quad in \quad L^\infty(Q_T)$$

and

$$\frac{y^{u^*+\varepsilon v} - y^{u^*}}{\varepsilon} \to f \quad in \quad L^\infty(Q_T),$$

as $\varepsilon \to 0+$, where $f = f(x,t)$ is the solution to

$$\begin{cases} \partial_t f - d\Delta f = a(x) f - 2k(x) y^{u^*} f - \chi_\omega(x) u^* f - \chi_\omega(x) v y^{u^*}, & (x,t) \in Q_T, \\ \partial_\nu f(x,t) = 0, & (x,t) \in \Sigma_T, \\ f(x,0) = 0. & x \in \Omega. \end{cases} \tag{8}$$

Returning to the proof of the theorem, passing to the limit in relation (7) and taking into consideration the results above, we obtain that:

$$\int_0^T \int_\omega u^*(x,t) f(x,t) dxdt + \int_0^T \int_\omega v(x,t) y^{u^*}(x,t) dxdt \leq 0. \tag{9}$$

We multiply the parabolic equation in (8) by p and integrate on Q_T. We get that:

$$\int_\Omega [p(x,T)f(x,T) - p(x,0)f(x,0)]dx - \int_0^T\int_\Omega f\partial_t p dxdt - \int_0^T\int_\Omega df\Delta p dxdt =$$

$$= \int_0^T a(x)pf dxdt - \int_0^T\int_\Omega 2k(x)y^{u^*}fp dxdt - \int_0^T\int_\omega u^* pf dxdt - \int_0^T\int_\omega pvy^{u^*}dxdt.$$

We using the fact that p is the solution of the problem (5) and we obtain that

$$-\int_0^T\int_\omega fu^*\,dxdt = -\int_0^T\int_\omega pvy^{u^*}\,dxdt. \tag{10}$$

From (9) and (10) we get that

$$\int_0^T\int_\omega v(x,t)y^{u^*}(x,t)(1+p(x,t))dxdt \le 0, \quad \text{for any } v \in L^\infty(\omega\times(0,T)),$$

such that $u^*+\varepsilon v \in K$, for sufficiently small $\varepsilon > 0$ (we have used the positivity of y^{u^*} in Q_T). So, the optimal control satisfies (6).

Next we remind an iterative method to improve the position of the support of the harvesting effort obtained in [8] for the model without logistic term. So, in the follows we will ignore the logistic process, i.e., we will take the case $k \equiv 0$. Let us consider $\omega_0 \subset \Omega$, where $\omega_0 \subset \Omega$ is a nonempty domain with sufficiently smooth boundary $\partial\omega_0$ and such that $\Omega \setminus \overline{\omega}_0$ is a domain. We denote by $\mathcal{O}$ the set

$$\mathcal{O} = \{\omega_0 + V \subset \Omega;\ V \in \mathbf{R}^N\}.$$

For any arbitrary but fixed $\omega \in \mathcal{O}$, we denote by (u^*_ω, y^*_ω) an optimal pair for problem (2). The optimal control problem to be investigated is:

$$Maximize_{\omega\in\mathcal{O}} \int_0^T\int_\omega u^*_\omega(x,t)y^*_\omega(x,t)dxdt, \tag{11}$$

where $y^*_\omega = y^*_\omega(x,t)$ is the solution to problem

$$\begin{cases} \partial_t y(x,t) - d\Delta y(x,t) = a(x)y(x,t) - \chi_\omega(x)u^*_\omega(x,t)y(x,t), & (x,t)\in Q_T,\\ \partial_\nu y(x,t) = 0, & (x,t)\in\Sigma_T,\\ y(x,0) = y_0(x), & x\in\Omega \end{cases} \tag{12}$$

In this case, the adjoint system is

$$\begin{cases} \partial_t p + d\Delta p = -a(x)p + \chi_\omega(x)u^*_\omega(1+p), & (x,t) \in Q_T, \\ \partial_\nu p(x,t) = 0, & (x,t) \in \Sigma_T, \\ p(x,T) = 0, & x \in \Omega \end{cases} \tag{13}$$

and the optimal control is given by

$$u^*_\omega(x,t) = \begin{cases} 0, & 1 + p_\omega(x,t) < 0 \\ L, & 1 + p_\omega(x,t) > 0 \end{cases} \tag{14}$$

a.e. $(x,t) \in \omega \times (0,T)$, where $p_\omega = p_\omega(x,t)$ is the solution to (13).
By (13) and (14) we get that p_ω is the solution to

$$\begin{cases} \partial_t p + d\Delta p = -a(x)p + \chi_\omega(x)L(1+p)^+, & (x,t) \in Q_T, \\ \partial_\nu p(x,t) = 0, & (x,t) \in \Sigma_T, \\ p(x,T) = 0, & x \in \Omega. \end{cases} \tag{15}$$

Multiplying (12) by p_ω and multiplying (13) by y^*_ω, and both integrating on Q_T we obtain:

$$\int_0^T \int_\omega u^*_\omega(x,t) y^*_\omega(x,t) dx dt = -\int_\Omega y_0(x) p_\omega(x,0) dx.$$

In conclusion our problem of optimal harvesting becomes a problem of minimizing another functional with respect to the positions of ω.
Let us denote

$$J^\omega = \int_\Omega y_0(x) p_\omega(x,0) dx,$$

where p_ω is the solution to (15).
Hence the minimization problem to be investigated is

$$Minimize_{\omega \in \mathcal{O}} J^\omega. \tag{16}$$

For every $V \in \mathbf{R}^n$, consider the derivative of J^ω with respect to translations. Actually

$$J^{\omega+\varepsilon V} - J^\omega = \int_\Omega (p_{\omega+\varepsilon V}(x,0) - p_\omega(x,0)) y_0(x) dx$$

and multiplying with $\frac{1}{\varepsilon}$ we have

$$\frac{1}{\varepsilon}[J^{\omega+\varepsilon V} - J^\omega] = \int_\Omega \frac{p_{\omega+\varepsilon V}(x,0) - p_\omega(x,0)}{\varepsilon} y_0(x) dx$$

For $\varepsilon \to 0+$ we obtain that

$$dJ^{\omega}(V) = \int_{\Omega} z(x,0)y_0(x)dx,$$

where $z = z(x,t)$ is the solution of the following boundary value problem:

$$\begin{cases} \partial_t z + d\Delta z = -a(x)z + L\chi_{\omega} z \partial h(1 + p_{\omega}(x,t)) + Lm_{p_{\omega}(t)}, & (x,t) \in Q_T, \\ \partial_{\nu} z(x,t) = 0, & (x,t) \in \Sigma_T, \\ z(x,T) = 0, & x \in \Omega \end{cases} \tag{17}$$

Here $h(r) = r^+$,

$$\partial h(r) = \begin{cases} 1, & r > 0 \\ I, & r = 0 \\ 0, & r < 0 \end{cases}$$

where $I = [0,1]$, and

$$m_{p_{\omega}(t)}(\varphi) = \int_{\partial\omega} (1 + p_{\omega}(x,t))^+ \varphi(x) V \cdot \nu(x) d\sigma, \text{ for any } \varphi \in H^1(\Omega)$$

where $\nu(x)$ is the outward normal versor at x to $\partial\omega$, outward with respect to $\Omega \setminus \omega$. We need to evaluate the form of the directional derivative for our functional. We recall the following result obtained in [8]:

Theorem 3 *For any $\omega \in \mathcal{O}$ and for any $V \in \mathbf{R}^N$,*

$$dJ^{\omega}(V) = -LV \cdot \int_0^T \int_{\partial\omega} (1 + p_{\omega}(x,t))^+ g_{\omega}(x,t)\nu(x) d\sigma dt,$$

where p_{ω} is the solution for (15) and $g_{\omega} = g_{\omega}(x,t)$ is the solution for the following boundary value problem:

$$\begin{cases} \partial_t g - d\Delta g = a(x)g - L\chi_{\omega} g \partial h(1 + p_{\omega}(x,t)), & (x,t) \in Q_T, \\ \partial_{\nu} g(x,t) = 0, & (x,t) \in \Sigma_T, \\ g(x,0) = y_0(x), & x \in \Omega \end{cases} \tag{18}$$

(For basic properties of the solution to such a problem we refer to [10]).
Proof. We multiply equation (17) with g_{ω} and we integrate on Q_T. This yields

$$\int_0^T \int_{\Omega} g_{\omega}(\partial_t z + d\Delta z + a(x)z) dx dt = L \int_0^T \int_{\omega} z(x,t)\xi(x,t) g_{\omega}(x,t) dx dt +$$

$$+L\int_0^T\int_{\partial\omega}(1+p_\omega)^+g_\omega(x,t)V\cdot\nu(x)d\sigma dt$$

where $\xi(x,t)\in\partial h(1+p_\omega(x,t))$ a.e. $(x,t)\in\omega\times(0,T)$.
Integrating by parts and using the fact that $z(x,T)=0$ and $g_\omega(x,0)=y_0(x)$ we obtain that

$$-\int_\Omega y_0(x)z(x,0)dx-\int_0^T\int_\Omega z[\partial_t g_\omega-d\Delta g_\omega-a(x)g_\omega]dxdt$$

$$=L\int_0^T\int_\omega z\xi(x,t)g_\omega(x,t)dxdt+L\int_0^T\int_{\partial\omega}(1+p_\omega)^+g_\omega(x,t)V\cdot\nu(x)d\sigma dt$$

and from (18) we get that:

$$-\int_\Omega y_0(x)z(x,0)dx=L\int_0^T\int_{\partial\omega}(1+p_\omega)^+g_\omega(x,t)V\cdot\nu(x)d\sigma dt.$$

The directional derivative of J^ω will be of the form

$$dJ^\omega(V)=-L\int_0^T\int_{\partial\omega}(1+p_\omega(x,t))^+g_\omega(x,t)V\cdot\nu(x)d\sigma dt$$

and we get the conclusion of the theorem.

3 A numerical algorithm

From Theorem 3 we derive the following conceptual iterative algorithm, based on a gradient method, to improve the position (translation) of $\omega\in\mathcal{O}$ in order to obtain a smaller value for J^ω.

Step 0: set $k:=0$, $J^{(0)}:=10^6$.
choose $\omega^{(0)}$ the initial positon of ω.

Step 1: compute $p^{(k+1)}$ the solution of the adjoint problem (15) corresponding to $\omega^{(k)}$.
compute $J^{(k+1)}=\int\limits_\Omega y_0(x)p^{(k+1)}(x,0)dx$.

Step 2: if $\left|J^{(k+1)}-J^{(k)}\right|<\varepsilon_1$ or $J^{(k+1)}\geq J^{(k)}$
then **STOP** ($\omega^{(k)}$ is the optimal position of ω)
else go to **Step 3**.

Step 3: compute $g^{(k+1)}$ the solution of problem (18) corresponding to

$\omega^{(k)}$ and $p^{(k+1)}$.

Step 4: compute

$$V := -\int_0^T \int_{\partial\omega^{(k)}} \left(1 + p^{(k+1)}(x,t)\right)^+ g^{(k+1)}(x,t)\nu(x)d\sigma dt$$

if $|V| < \varepsilon_2$
then **STOP** ($\omega^{(k)}$ is the optimal position of ω)
else go to **Step 5**.

Step 5: compute the new position of ω
$\omega^{(k+1)} := \rho V + \omega^{(k)}$;

Step 6: if $\omega^{(k+1)} = \omega^{(k)}$
then **STOP** ($\omega^{(k)}$ is the optimal position of ω)
else $k := k + 1$;
go to **Step 1**.

In Step 5, $\rho > 0$ is a given parameter (the gradient steplength), and $\varepsilon_1 > 0$ in Step 2 and $\varepsilon_2 > 0$ in Step 4 are prescribed convergence parameters.
The conceptual iterative algorithm, used to improve the position of ω in order to obtain a smaller value for J^ω, is a descent method. For more information about gradient (descent) methods, see [9], Section 2.3. The steplength ρ from Step 5 is variable from an iteration to the next one. To fit it we have used Armijo method (see [7] for more details).

4 Numerical tests

In order to simplify the discretization formulae for the numerical tests we have considered Ω and $\omega^{(0)}$ to be squares with the sides parallel with Ox_1 and Ox_2 axes (the space variable is $x = (x_1, x_2)$). Let $\Omega = (0,1) \times (0,1)$ and the length side of ω is equal with 0.2. We introduce equidistant discretization nodes for both axis corresponding to Ω. The interval $[0, T]$ is also discretized by equidistant nodes. The parabolic system from Step 1 is approximated by a finite difference method, descending with respect to time levels. An implicit scheme is used. The resulting algebraic linear system is solved by Gaussian elimination. The parabolic system from Step 3 is approximated also using a finite difference method, but ascending with respect to time levels. Integrals from Step 1 and Step 4 are numerical computed using Simpson's method corresponding to the discrete grid. In all following figures the square drawn with solid line represent the initial position of ω, and the

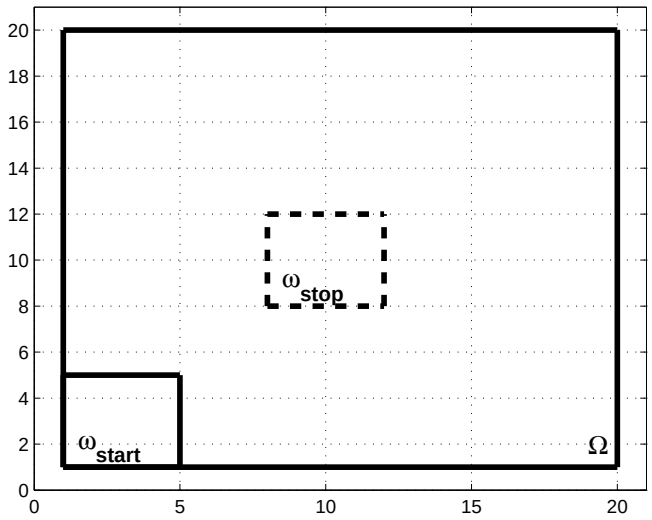

Figure 1. START/STOP position of ω

square drawn with dashed line is the improved position of ω.

Test 1. We consider the natural growth rate of the population $a(x_1, x_2) = 5, (x_1, x_2) \in \Omega$, $d = 1$, and final time $T = 1$. We take the space discretization step $\Delta x_1 = \Delta x_2 = 0.05$, and the time discretization step $\Delta t = 0.025$. The nodes along both axes Ox_1 and Ox_2 are numbered from 1 to 20. The left-down corner of Ω is numbered as $(1, 1)$ while the right-up corner is numbered as $(20, 20)$. For the convergence tests we consider $\varepsilon_1 = \varepsilon_2 = 0.001$. We start with $\omega^{(0)}$ which has the left-down corner at node $(1, 1)$ and the MATLAB program corresponding to the above algorithm gives after 5 iterations the optimal ω which has the left-down corner at node $(8, 8)$. The convergence was obtained by the test in Step 2. The initial and the optimal position of ω are shown in Figure 1. The corresponding graph of the optimal state $y(x_1, x_2, t)$ for $t = 1$ is given in Figure 2.

Test 2. The results obtained using the same input data from example 1, except the initial position of ω, are shown in Figure 3. We start with $\omega^{(0)}$ which has the left-down corner at node $(16, 10)$ and the MATLAB program corresponding to the above algorithm gives after 5 iterations the optimal

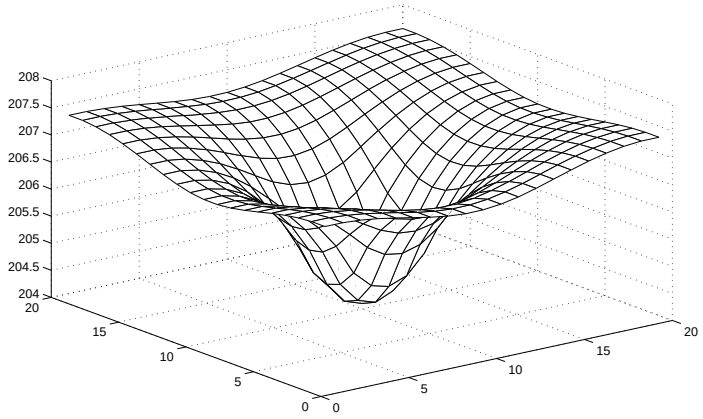

Figure 2. The optimal state y for $t = 1$

ω which has the left-down corner at the node $(7, 8)$. The convergence was obtained by the test in Step 4.

Test 3. We consider $a(x_1, x_2) = x_2 - x_1, (x_1, x_2) \in \Omega$, $d = 1$, and final time $T = 1$. We take the space discretization step $\Delta x_1 = \Delta x_2 = 0.05$, and the time discretization step $\Delta t = 0.05$ since the finite difference method used is implicit. We start with $\omega^{(0)}$ which has the left-down corner at node $(16, 16)$ and the MATLAB program corresponding to the above algorithm gives after 4 iterations the optimal ω which has the left-down corner at the node $(7, 9)$ (see Figure 4). The convergence was obtained by the test in Step 4.

Let us point out that the final position of ω given by the computer program is central with respect to Ω no matter the starting position $\omega^{(0)}$. This is in accordance with a more general theoretical result obtained in [2].

Test 4. For the natural growth rate of the population $a(x_1, x_2) = x_2 sin(x_1)$, $(x_1, x_2) \in \Omega$, the optimal position of ω is no more central with respect to Ω. The left-down corner of $\omega^{(0)}$ is $(4, 7)$ and the left-hand corner of the final ω is $(2, 4)$ and it is obtained after 4 iterations (see Figure 5). The convergence was obtained by the test in Step 4.

Let us point out that the algorithm is fast according to the number of iterations.

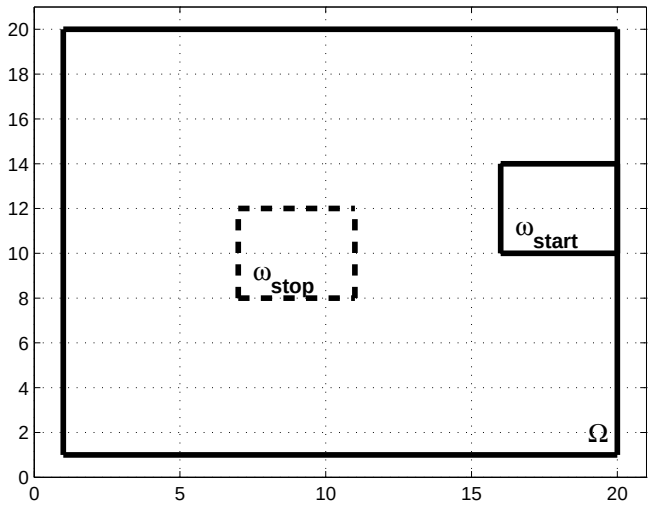

Figure 3. START/STOP position of ω

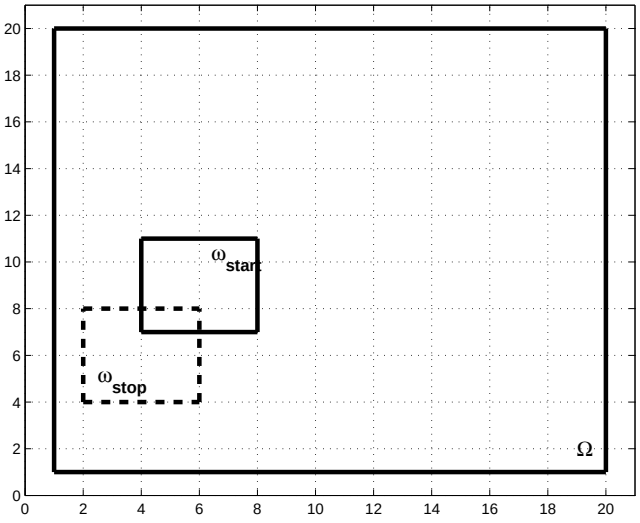

Figure 4. START/STOP position of ω

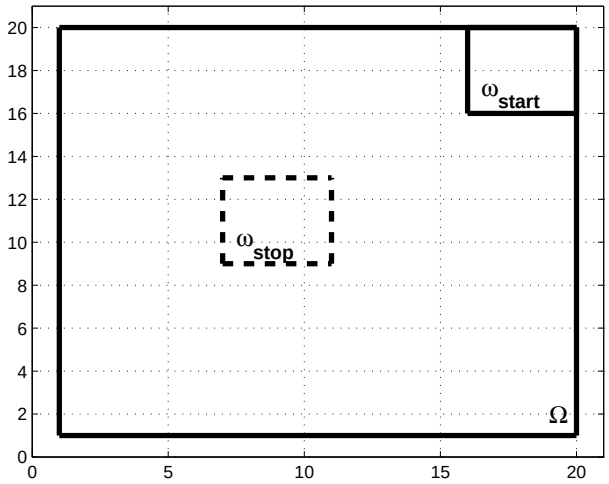

Figure 5. START/STOP position of ω

Acknowledgments

This work was supported by a grant of the Ministry of National Education, CNCS-UEFISCDI, project number PN-II-ID-PCE-2012-4-0270 (68/2. 09.2013): "Optimal Control and Stabilization of Nonlinear Parabolic Systems with State Constraints. Applications in Life Sciences and Economics".

References

[1] L.-I. Aniţa, S. Aniţa. Note on the stabilization of a reaction-diffusion model in epidemiology. *Nonlin. Anal.; Real World Appl.* 6:537–544, 2005.

[2] L.-I. Aniţa, S. Aniţa, V. Arnăutu. Internal null stabilization for some diffusive models of population dynamics. *Appl. Math. Comput.* 219:10231–10244, 2013.

[3] S. Aniţa. *Analysis and control of age-dependent population dynamics.* Kluwer Academic, Dordrecht, 2000.

[4] S. Aniţa, V. Arnăutu, V. Capasso. *An introduction to optimal control problems in life sciences and economics. From mathematical models to numerical simulation with Matlab.* Birkhauser, Boston, 2011.

[5] S. Aniţa, V. Arnăutu, S. Dodea. Feedback stabilization for a reaction-diffusion system with nonlocal reaction term. *Numer. Funct. Anal. Optimiz.* 32:1-19, 2011.

[6] S. Aniţa, J. Casas, C. Suppo. Impulsive spatial control of invading pests by generalist predators. *Math. Med. Biol.* 31(3):284–301, 2014.

[7] L. Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific J. Math.* 16:1–3, 1966.

[8] V. Arnăutu, A.-M. Moşneagu. Optimal control and stabilization for some Fisher–like models. *Numer. Funct. Anal. Optimiz.* 36 (2015), pp. 567-589.

[9] V. Arnăutu, P. Neittaanmäki. *Optimal control from theory to computer programs.* Kluwer Acad. Publ., Dordrecht, 2003.

[10] V. Barbu. *Analysis and control of nonlinear infinite dimensional systems.* Academic Press, San Diego, 1993.

[11] V. Barbu. *Mathematical methods in optimization of differential systems* Kluwer Acad. Publ., Dordrecht, 1994.

[12] R. A. Fisher. The wave of advance of advantageous genes. *Ann. Engen.* 7:355-369.

[13] A. Henrot, M. Pierre. *Variation et optimisation de formes. Une analyse géométrique*. Springer, Berlin, 2005.

[14] J.D. Murray. *Mathematical biology II. Spatial Models and Biomedical Applications, 3rd edition.* Springer-Verlag, New York, 2003.

[15] A. Okubo. *Diffusion and ecological problems: mathematical models*. Springer-Verlag, Berlin, 1980.

[16] A. Okubo. Dynamical aspects of anumail grouping: swarms, schools, flocks and herbs.*Adv. Biophys.* 22:1–94, 1986.

Ann. Acad. Rom. Sci.
Ser. Math. Appl.
ISSN 2066-6594 Vol. 7, No. 1/2015

SECOND-ORDER ANALYSIS OF A BOUNDARY CONTROL PROBLEM FOR THE VISCOUS CAHN–HILLIARD EQUATION WITH DYNAMIC BOUNDARY CONDITION*

Pierluigi Colli† Mohammad Hassan Farshbaf-Shaker‡
Gianni Gilardi§ Jürgen Sprekels¶

Dedicated to the memory of Prof. Dr. Viorel Arnăutu

Abstract

In this paper we establish second-order sufficient optimality conditions for a boundary control problem that has been introduced and studied by three of the authors in the preprint arXiv:1407.3916. This control problem regards the viscous Cahn–Hilliard equation with possibly singular potentials and dynamic boundary conditions.

MSC: 49J20 (Primary), 49J50, 49K20 (Secondary)

*Accepted for publication on December 10th, 2014

†pierluigi.colli@unipv.it Dipartimento di Matematica "F. Casorati", Università di Pavia, via Ferrata 1, 27100 Pavia, Italy

‡Hassan.Farshbaf-Shaker@wias-berlin.de Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstrasse 39, 10117 Berlin, Germany

§gianni.gilardi@unipv.it Dipartimento di Matematica "F. Casorati", Università di Pavia, via Ferrata 1, 27100 Pavia, Italy

¶sprekels@wias-berlin.de Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstrasse 39, 10117 Berlin and Department of Mathematics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

keywords: Cahn–Hilliard equation, dynamic boundary conditions, phase separation, singular potentials, optimal control, first and second order optimality conditions, adjoint state system.

1 Introduction

This paper deals with second-order optimality conditions of a special boundary control problem for the viscous Cahn–Hilliard equation with dynamic boundary conditions. It continues the work [2] by three of the present authors in which the first-order necessary conditions of optimality were derived. For the work of other authors concerning the optimal control of Cahn–Hilliard systems, we refer the reader to the references given in [2].

Crucial contributions in [2] were the derivation of the adjoint problem, whose form turned out to be nonstandard, and an existence result for its solutions. As is well known, first-order conditions are in the case of nonlinear equations usually not sufficient for optimality. Also, second-order sufficient optimality conditions for nonlinear optimal control problems are essential both in the numerical analysis and for the construction of reliable optimization algorithms. For instance, the strong convergence of optimal controls and states for numerical discretizations of the problem rests heavily on the availability of second-order sufficient optimality conditions; furthermore, one can show that numerical algorithms such as SQP methods are locally convergent if second-order sufficient optimality conditions hold true. For a general discussion of second-order sufficient conditions for elliptic and parabolic control problems, we refer the reader to [6] and references therein; for the case of control problems involving phase field models, we refer to, e. g., [3, 5].

In this paper, we aim to establish second-order sufficient optimality conditions for the boundary control problem studied in [2]. To this end, we assume that an open, bounded and connected set $\Omega \subset \mathbb{R}^3$, with smooth boundary Γ and unit outward normal $\mathbf{n}$, and some final time $T > 0$ are given, and we set $Q := \Omega \times (0,T)$ and $\Sigma := \Gamma \times (0,T)$. Moreover, we denote by Δ_Γ, ∇_Γ, $\partial_\mathbf{n}$, the Laplace–Beltrami operator, the surface gradient, and the outward normal derivative on Γ, in this order. We make the following general assumptions:

(A1) There are given nonnegative constants $b_Q, b_\Sigma, b_\Omega, b_\Gamma, b_0$, which do not all vanish, functions $z_Q \in L^2(Q)$, $z_\Sigma \in L^2(\Sigma)$, $z_\Omega \in L^2(\Omega)$, $z_\Gamma \in L^2(\Gamma)$, as well as a constant $M_0 > 0$ and functions $u_{\Gamma,\min} \in L^\infty(\Sigma)$ and $u_{\Gamma,\max} \in L^\infty(\Sigma)$ with $u_{\Gamma,\min} \le u_{\Gamma,\max}$ a. e. in Σ.

(A2) There are given constants $-\infty \le r_- < 0 < r_+ \le +\infty$ and two functions $f, f_\Gamma : (r_-, r_+) \to [0, +\infty)$ such that the following holds:

$$f, f_\Gamma \in C^4(r_-, r_+), \quad f(0) = f_\Gamma(0) = 0, \tag{1}$$

$$f'' \text{ and } f_\Gamma'' \text{ are bounded from below}, \tag{2}$$

$$\lim_{r \searrow r_-} f'(r) = \lim_{r \searrow r_-} f_\Gamma'(r) = -\infty \quad \text{and} \quad \lim_{r \nearrow r_+} f'(r) = \lim_{r \nearrow r_+} f_\Gamma'(r) = +\infty\,, \tag{3}$$

$$|f'(r)| \le \eta\, |f_\Gamma'(r)| + C \quad \text{for some } \eta,\, C > 0 \text{ and every } r \in (r_-, r_+). \tag{4}$$

In fact, (1) is fully used only in the last part of the paper, and many of our results hold under a weaker assumption. We also note that the conditions (1)–(4) allow for the possibility of splitting f' in (3) in the form $f' = \beta + \pi$, where β is a monotone function that diverges at $r_\pm$ and π is a perturbation having a bounded derivative. Since the same is true for f_Γ, the general assumptions of [1] are satisfied. Typical and important examples for f and f_Γ are the classical regular potential $f_{\rm reg}$ and the logarithmic double-well potential $f_{\rm log}$ given by

$$f_{\rm reg}(r) = \frac{1}{4}(r^2 - 1)^2\,, \quad r \in \mathbb{R} \tag{5}$$

$$f_{\rm log}(r) = ((1+r)\ln(1+r) + (1-r)\ln(1-r)) - cr^2\,, \quad r \in (-1, 1), \tag{6}$$

where in the latter case we assume that $c > 0$ is so large that $f_{\rm log}$ is non-convex.

With the above assumptions, we consider the following tracking type optimal boundary control problem:
(**CP**) Minimize

$$\begin{aligned} \mathcal{J}(y, y_\Gamma, u_\Gamma) := {} & \frac{b_Q}{2} \|y - z_Q\|^2_{L^2(Q)} + \frac{b_\Sigma}{2} \|y_\Gamma - z_\Sigma\|^2_{L^2(\Sigma)} \\ & + \frac{b_\Omega}{2} \|y(T) - z_\Omega\|^2_{L^2(\Omega)} + \frac{b_\Gamma}{2} \|y_\Gamma(T) - z_\Gamma\|^2_{L^2(\Gamma)} + \frac{b_0}{2} \|u_\Gamma\|^2_{L^2(\Sigma)} \end{aligned} \tag{7}$$

subject to the control constraint

$$\begin{aligned} u_\Gamma \in \mathcal{U}_{ad} := {} & \big\{ v_\Gamma \in H^1(0, T; L^2(\Gamma)) \cap L^\infty(\Sigma) : \\ & u_{\Gamma,\min} \le v_\Gamma \le u_{\Gamma,\max} \text{ a.e. on } \Sigma,\ \|\partial_t v_\Gamma\|_2 \le M_0 \big\} \end{aligned} \tag{8}$$

and to the Cahn–Hilliard equation with nonlinear dynamic boundary con-

ditions as the state system,

$$\partial_t y - \Delta w = 0 \quad \text{in } Q, \tag{9}$$

$$w = \partial_t y - \Delta y + f'(y) \quad \text{in } Q, \tag{10}$$

$$\partial_{\mathbf{n}} w = 0 \quad \text{on } \Sigma, \tag{11}$$

$$y_\Gamma = y_{|\Gamma} \quad \text{on } \Sigma, \tag{12}$$

$$\partial_t y_\Gamma + \partial_{\mathbf{n}} y - \Delta_\Gamma y_\Gamma + f'_\Gamma(y_\Gamma) = u_\Gamma \quad \text{on } \Sigma, \tag{13}$$

$$y(\cdot,0) = y_0 \quad \text{in } \Omega, \quad y_\Gamma(\cdot,0) = y_{0\Gamma} \quad \text{on } \Gamma. \tag{14}$$

Here, and throughout this paper, we generally assume that the admissible set $\mathcal{U}_{ad}$ is nonempty. Moreover, we postulate:

(A3) $y_0 \in H^2(\Omega)$, $y_{0\Gamma} := y_{0|\Gamma} \in H^2(\Gamma)$, and it holds (notice that $y_0 \in C^0(\overline{\Omega})$)

$$r_- < y_0 < r_+ \quad \text{in } \overline{\Omega}. \tag{15}$$

We remark at this place that in [1] the additional assumption $\partial_{\mathbf{n}} y_0 = 0$ was made; this postulate is however unnecessary for the results of [1] to hold, since it is nowhere used in the proofs.

The system (9)–(14) is an initial-boundary value problem with nonlinear dynamic boundary condition for a Cahn–Hilliard equation. In this connection, the unknown y usually stands for the order parameter of an isothermal phase transition, and w denotes the chemical potential of the system.

Our paper is organized as follows: in Section 2, we provide and collect some results proved in [2, 1] concerning the state system, and we study a certain linear counterpart thereof that will be employed repeatedly in the later analysis. In Section 3, the existence of the second-order Fréchet derivative of the control-to-state mapping will be shown. Section 4 then brings the derivation of the second-order sufficient condition of optimality.

In order to simplify notation, we will in the following write y_Γ for the trace $y_{|\Gamma}$ of a function $y \in H^1(\Omega)$ on Γ, and we introduce the abbreviations

$$\begin{aligned}
&V := H^1(\Omega), \ \ H := L^2(\Omega), \ \ V_\Gamma := H^1(\Gamma), \ \ H_\Gamma := L^2(\Gamma), \ \ \mathcal{H} := H \times H_\Gamma, \\
&\mathcal{V} := \{(v, v_\Gamma) \in V \times V_\Gamma : \ v_\Gamma = v_{|\Gamma}\}, \ \ \mathcal{G} := H^2(\Omega) \times H^2(\Gamma), \\
&\mathcal{X} := H^1(0,T;H_\Gamma) \cap L^\infty(\Sigma), \ \ \mathcal{Y} := H^1(0,T;\mathcal{H}) \cap L^\infty(0,T;\mathcal{V}),
\end{aligned} \tag{16}$$

and endow these spaces with their natural norms. Moreover, for the generic Banach space X we denote by X^* its dual space and by $\|\cdot\|_X$ its norm. Furthermore, the symbol $\langle\cdot\,,\cdot\rangle$ stands for the duality pairing between the spaces V^* and V, where it is understood that H is embedded in V^* in the usual way, i. e., such that we have $\langle u,v\rangle = (u,v)$ for every $u \in H$ and $v \in V$ with the standard inner product $(\cdot\,,\cdot)$ of H. Finally, for $u \in V^*$ and $v \in L^1(0,T;V^*)$ we define their generalized mean values $u^\Omega \in \mathbb{R}$ and $v^\Omega \in L^1(0,T)$, respectively, by setting

$$u^\Omega := \frac{1}{|\Omega|}\,\langle u,1\rangle \quad \text{and} \quad v^\Omega(t) := (v(t))^\Omega \quad \text{for a. e. } t \in (0,T), \tag{17}$$

where $|\Omega|$ stands for the Lebesgue measure of Ω.

During the course of our analysis, we will make repeated use of the elementary Young's inequality

$$ab \le \delta a^2 + \frac{1}{4\delta}\,b^2 \quad \text{for every } a,b \ge 0 \text{ and } \delta > 0, \tag{18}$$

of Hölder's inequality, and of Poincaré's inequality

$$\|v\|_V^2 \le \widehat{C}\big(\|\nabla v\|_H^2 + |v^\Omega|^2\big) \quad \text{for every } v \in V, \tag{19}$$

where $\widehat{C} > 0$ depends only on Ω.

Next, we recall a tool that is commonly used in the context of problems related to the Cahn–Hilliard equation. We define

$$\operatorname{dom}\mathcal{N} := \{v_* \in V^* :\ v_*^\Omega = 0\} \quad \text{and} \quad \mathcal{N} : \operatorname{dom}\mathcal{N} \to \{v \in V :\ v^\Omega = 0\} \tag{20}$$

by setting, for $v_* \in \operatorname{dom}\mathcal{N}$,

$$\mathcal{N}v_* \in V,\ \ (\mathcal{N}v_*)^\Omega = 0,\ \text{and}\ \int_\Omega \nabla\mathcal{N}v_* \cdot \nabla z\,dx = \langle v_*, z\rangle \quad \forall\, z \in V\,. \tag{21}$$

That is, $v = \mathcal{N}v_*$ is the unique weak solution with $v^\Omega = 0$ to the Neumann problem for $-\Delta$ with datum v_*. Indeed, if $v_* \in H$, then the above variational equation means that $-\Delta\mathcal{N}v_* = v_*$ in Ω and $\partial_{\mathbf{n}}\mathcal{N}v_* = 0$ on Γ. Moreover, we have

$$\langle u_*, \mathcal{N}v_*\rangle = \langle v_*, \mathcal{N}u_*\rangle = \int_\Omega (\nabla\mathcal{N}u_*)\cdot(\nabla\mathcal{N}v_*)\,dx \quad \forall\, u_*, v_* \in \operatorname{dom}\mathcal{N}, \tag{22}$$

whence also, for every $v_* \in H^1(0,T;V^*)$ satisfying $(v_*)^\Omega = 0$ a.e. in $(0,T)$,

$$2\langle \partial_t v_*(t), \mathcal{N}v_*(t)\rangle = \frac{d}{dt}\,\|v_*(t)\|_*^2 \quad \text{for a.a. } t \in (0,T), \tag{23}$$

where we set $\|v_*\|_*^2 := \displaystyle\int_\Omega |\nabla\mathcal{N}v_*|^2\,dx$ for every $v_* \in V^*$.

2 The state equation

At first, we specify our notion of solution to the state system (9)–(14).

Definition 1. Suppose that the general assumptions **(A1)**–**(A3)** are fulfilled, and let $u_\Gamma \in \mathcal{X}$ be given. By a *solution* to (9)–(14) we mean a triple (y, y_Γ, w) that satisfies

$$y \in W^{1,\infty}(0,T;H) \cap H^1(0,T;V) \cap L^\infty(0,T;H^2(\Omega)), \tag{24}$$

$$y_\Gamma \in W^{1,\infty}(0,T;H_\Gamma) \cap H^1(0,T;V_\Gamma) \cap L^\infty(0,T;H^2(\Gamma)), \tag{25}$$

$$y_\Gamma(t) = y(t)_{|\Gamma} \quad \text{for a. a. } t \in (0,T), \tag{26}$$

$$r_- < \inf_Q \operatorname{ess} y \le \sup_Q \operatorname{ess} y < r_+ , \quad r_- < \inf_\Sigma \operatorname{ess} y_\Gamma \le \sup_\Sigma \operatorname{ess} y_\Gamma < r_+ , \tag{27}$$

$$w \in L^\infty(0,T;H^2(\Omega)), \tag{28}$$

as well as, for almost every $t \in (0,T)$, the variational equations

$$\int_\Omega \partial_t y(t)\, v \, dx + \int_\Omega \nabla w(t) \cdot \nabla v \, dx = 0 , \tag{29}$$

$$\begin{aligned} \int_\Omega w(t)\, v \, dx &= \int_\Omega \partial_t y(t)\, v \, dx + \int_\Gamma \partial_t y_\Gamma(t)\, v_\Gamma \, d\Gamma + \int_\Omega \nabla y(t) \cdot \nabla v \, dx \\ &+ \int_\Gamma \nabla_\Gamma y_\Gamma(t) \cdot \nabla_\Gamma v_\Gamma \, d\Gamma + \int_\Omega f'(y(t))\, v \, dx \\ &+ \int_\Gamma \big(f_\Gamma'(y_\Gamma(t)) - u_\Gamma(t) \big)\, v_\Gamma \, d\Gamma, \end{aligned} \tag{30}$$

for every $v \in V$ and every $(v, v_\Gamma) \in \mathcal{V}$, respectively, and the Cauchy condition

$$y(0) = y_0 , \quad y_\Gamma(0) = y_{0\Gamma} . \tag{31}$$

Remark 1. It is worth noting that (recall the notation (17))

$$\begin{aligned} &(\partial_t y(t))^\Omega = 0 \ \text{ for a. a. } \ t \in (0,T) \ \text{ and } \ y(t)^\Omega = m_0 \ \text{ for every } \ t \in [0,T], \\ &\text{where } \ m_0 = (y_0)^\Omega \ \text{ is the mean value of } y_0, \end{aligned} \tag{32}$$

as usual for the Cahn–Hilliard equation.

Now recall that $\mathcal{U}_{ad}$ is a convex, closed, and bounded subset of the Banach space $\mathcal{X}$ and thus contained in some bounded open ball in $\mathcal{X}$. For

convenience, we fix such a ball once and for all, noting that any other such ball could be used instead. The next assumption is thus rather a denotation:

(A4) The set $\mathcal{U}$ is some open ball in $\mathcal{X}$ that contains $\mathcal{U}_{ad}$ and satisfies

$$\|u_\Gamma\|_{H^1(0,T;L^2(\Gamma))} + \|u_\Gamma\|_{L^\infty(\Sigma)} \le R \quad \forall\, u_\Gamma \in \mathcal{U}, \tag{33}$$

where $R > 0$ is a fixed given constant.

Concerning the well-posedness of the state system, we have the following result.

Theorem 1. *Suppose that the general hypotheses* **(A1)**–**(A4)** *are fulfilled. Then the state system* (9)–(14) *has for any* $u_\Gamma \in \mathcal{U}$ *a unique solution* (y, y_Γ, w) *in the sense of Definition 1. Moreover, there are constants* $K_1^* > 0$, $K_2^* > 0$, *and* $\widetilde{r}_-, \widetilde{r}_+ \in (r_-, r_+)$, *which depend only on* Ω, T, *the shape of the nonlinearities* f *and* f_Γ, *the initial datum* y_0, *and the constant* R, *such that the following holds:*
(i) *Whenever* (y, y_Γ, w) *is the solution to* (9)–(14) *associated with some* $u_\Gamma \in \mathcal{U}$, *then*

$$\|(y, y_\Gamma)\|_{W^{1,\infty}(0,T;\mathcal{H})\cap H^1(0,T;\mathcal{V})\cap L^\infty(0,T;\mathcal{G})} + \|w\|_{L^\infty(0,T;H^2(\Omega))} \le K_1^*, \tag{34}$$

$$\widetilde{r}_- \le y \le \widetilde{r}_+ \quad a.e.\ in\ Q, \qquad \widetilde{r}_- \le y_\Gamma \le \widetilde{r}_+ \quad a.e.\ on\ \Sigma. \tag{35}$$

(ii) *Whenever* $(y_i, y_{i,\Gamma}, w_i)$, $i = 1, 2$, *are the solutions to* (9)–(14) *associated with* $u_{i,\Gamma} \in \mathcal{U}$, $i = 1, 2$, *then*

$$\|(y_1, y_{1,\Gamma}) - (y_2, y_{2,\Gamma})\|_{H^1(0,T;\mathcal{H})\cap L^\infty(0,T;\mathcal{V})} \le K_2^* \, \|u_{1,\Gamma} - u_{2,\Gamma}\|_{L^2(\Sigma)}. \tag{36}$$

Proof. We may apply Theorems 2.2, 2.3, 2.4, 2.6 and Corollary 2.7 of [1] (where $\mathcal{V}$ has a slightly different meaning with respect to the present paper) to deduce that (i) holds true. Moreover, assertion (ii) is a consequence of [2, Lemma 4.1]. □

Remark 2. It follows from Theorem 1 that the control-to-state operator

$$\mathcal{S} : \mathcal{U} \to W^{1,\infty}(0,T;\mathcal{H}) \cap H^1(0,T;\mathcal{V}) \cap L^\infty(0,T;\mathcal{G}), \quad u_\Gamma \mapsto (y, y_\Gamma), \tag{37}$$

is well defined and Lipschitz continuous from $\mathcal{U}$, viewed as a subset of $L^2(\Sigma)$, into $\mathcal{Y}$. Moreover, owing to (34) and (35), we may assume (by possibly choosing a larger K_1^*) that for any $u_\Gamma \in \mathcal{U}$ the corresponding state $(y, y_\Gamma) = \mathcal{S}(u_\Gamma)$ satisfies

$$\max_{1\le i\le 4} \left(\|f^{(i)}(y)\|_{L^\infty(Q)} + \|f_\Gamma^{(i)}(y_\Gamma)\|_{L^\infty(\Sigma)} \right) \le K_1^*. \tag{38}$$

Next, in order to ensure the solvability of a number of linearized systems later in this paper, we introduce the linear initial-boundary value problem

$$\partial_t \chi - \Delta \mu = 0 \quad \text{in } Q, \tag{39}$$

$$\mu = \partial_t \chi - \Delta \chi + \lambda \chi + g \quad \text{in } Q, \tag{40}$$

$$\partial_{\mathbf{n}} \mu = 0 \quad \text{on } \Sigma, \tag{41}$$

$$\chi_\Gamma = \chi_{|\Gamma} \quad \text{on } \Sigma, \tag{42}$$

$$\partial_t \chi_\Gamma + \partial_{\mathbf{n}} \chi - \Delta_\Gamma \chi_\Gamma + \lambda_\Gamma \chi_\Gamma = g_\Gamma \quad \text{on } \Sigma, \tag{43}$$

$$\chi(0) = \chi_0 \quad \text{in } \Omega, \quad \chi_\Gamma(0) = \chi_{0_\Gamma} := \chi_{0|\Gamma} \quad \text{on } \Gamma, \tag{44}$$

and its variational counterpart, namely, for almost every $t \in (0,T)$,

$$\int_\Omega \partial_t \chi(t)\, v \, dx + \int_\Omega \nabla \mu(t) \cdot \nabla v \, dx = 0 \quad \text{for every } v \in V, \tag{45}$$

$$\begin{aligned} \int_\Omega \mu(t) v \, dx = {} & \int_\Omega \partial_t \chi(t)\, v \, dx + \int_\Gamma \partial_t \chi_\Gamma(t)\, v_\Gamma \, d\Gamma + \int_\Omega \nabla \chi(t) \cdot \nabla v \, dx \\ & + \int_\Gamma \nabla_\Gamma \chi_\Gamma(t) \cdot \nabla_\Gamma v_\Gamma \, d\Gamma + \int_\Omega \big(\lambda(t)\, \chi(t) + g(t)\big)\, v \, dx \\ & + \int_\Gamma \big(\lambda_\Gamma(t)\, \chi_\Gamma(t) - g_\Gamma(t)\big)\, v_\Gamma \, d\Gamma \quad \text{for every } (v, v_\Gamma) \in \mathcal{V}, \end{aligned} \tag{46}$$

together with the Cauchy condition

$$\chi(0) = \chi_0, \quad \chi_\Gamma(0) = \chi_{0_\Gamma}. \tag{47}$$

We have the following result.

Lemma 1. *Suppose that* $(g, g_\Gamma) \in H^1(0,T;\mathcal{H}) \cap (L^\infty(Q) \times L^\infty(\Sigma))$ *and* $(\lambda, \lambda_\Gamma) \in W^{1,\infty}(0,T;\mathcal{H}) \cap (L^\infty(Q) \times L^\infty(\Sigma))$ *are given, and let* $\chi_0 \in H^2(\Omega)$ *be such that* $\chi_{0_\Gamma} := \chi_{0|\Gamma} \in H^2(\Gamma)$. *Then the problem* (39)–(44) *has a unique solution in the sense that there is a unique triple* (χ, χ_Γ, μ) *that fulfills* (45)–(47) *and whose components satisfy the analogue of the regularity requirements* (24), (25), *and* (28), *respectively. Moreover, there exists a constant* $K_3^* > 0$, *which depends only on* Ω, T, $\|\lambda\|_{L^\infty(Q)}$, *and* $\|\lambda_\Gamma\|_{L^\infty(\Sigma)}$, *such that the following holds: whenever* $\chi_0 = 0$, *then*

$$\|(\chi, \chi_\Gamma)\|_{H^1(0,T;\mathcal{H}) \cap L^\infty(0,T;\mathcal{V})} \le K_3^* \, \|(g, g_\Gamma)\|_{L^2(0,T;\mathcal{H})}. \tag{48}$$

Proof. In the following, we denote by C_i, $i \in \mathbb{N}$, positive constants that only depend on the quantities mentioned in the assertion. First, we observe that the results concerning existence, uniqueness, and regularity follow from a direct application of [1, Cor. 2.5]. Now assume that $\chi_0 = 0$. Then we have $\chi^\Omega(t) = 0$ for almost every $t \in (0,T)$. We thus may choose in (45) $v = \mathcal{N}(\chi(t))$, and in (46) $v = -\chi(t)$. Adding the resulting equalities and integrating with respect to time, we arrive at the identity

$$\begin{aligned}\frac{1}{2}\big(\|\chi(t)\|_*^2 + \|\chi(t)\|_H^2 + \|\chi_\Gamma(t)\|_{H_\Gamma}^2\big) + \int_0^t\!\!\int_\Omega |\nabla\chi|^2\,dx\,ds + \int_0^t\!\!\int_\Gamma |\nabla_\Gamma\chi_\Gamma|^2\,d\Gamma\,ds \\ = \int_0^t\!\!\int_\Omega \big(-g - \lambda\,\chi\big)\,\chi\,dx\,ds + \int_0^t\!\!\int_\Gamma \big(g_\Gamma - \lambda_\Gamma\,\chi_\Gamma\big)\,\chi_\Gamma\,d\Gamma\,ds\end{aligned}$$

for all $t \in [0,T]$. Estimating the right-hand side with the help of Young's and Poincaré's inequalities, and applying Gronwall's lemma, we have that

$$\|(\chi,\chi_\Gamma)\|_{L^\infty(0,T;\mathcal{H})\cap L^2(0,T;\mathcal{V})} \le C_1\,\|(g,g_\Gamma)\|_{L^2(0,T;\mathcal{H})}. \tag{49}$$

Moreover, we may insert $v = \mathcal{N}(\partial_t\chi(t))$ in (45) and $v = -\partial_t\chi(t)$ in (46). Adding the resulting equations, integrating with respect to time, and using (21), we obtain the identity

$$\begin{aligned}&\int_0^t \|\partial_t\chi(s)\|_*^2\,ds + \int_0^t\!\!\int_\Omega |\partial_t\chi|^2\,dx\,ds + \int_0^t\!\!\int_\Gamma |\partial_t\chi_\Gamma|^2\,d\Gamma\,ds \\ &\quad + \frac{1}{2}\,(\|\nabla\chi(t)\|_H^2 + \|\nabla_\Gamma\chi_\Gamma(t)\|_{H_\Gamma}^2) \\ &= \int_0^t\!\!\int_\Omega (-g - \lambda\,\chi)\,\partial_t\chi\,dx\,ds + \int_0^t\!\!\int_\Gamma (g_\Gamma - \lambda_\Gamma\,\chi_\Gamma)\,\partial_t\chi_\Gamma\,d\Gamma\,ds\,.\end{aligned} \tag{50}$$

Invoking Young's inequality, we can easily infer from (49) and (50) the estimate

$$\|(\chi,\chi_\Gamma)\|_{H^1(0,T;\mathcal{H})\cap L^\infty(0,T;\mathcal{V})} \le C_2\,\|(g,g_\Gamma)\|_{L^2(0,T;\mathcal{H})}, \tag{51}$$

whence the assertion follows. □

3 Differentiability properties of the control-to-state mapping

The main objective in this section is to prove that the control-to-state mapping is twice continuously differentiable. We begin our analysis with the following result.

Theorem 2. *Suppose that* **(A1)**–**(A4)** *are fulfilled. Then the following holds true:*
(i) *The control-to-state mapping* $\mathcal{S}$ *is Fréchet differentiable in* $\mathcal{U}$ *as a mapping from* $\mathcal{U} \subset \mathcal{X}$ *to* $\mathcal{Y}$.
(ii) *Let* $u_\Gamma \in \mathcal{U}$, *and let* $(y, y_\Gamma) = \mathcal{S}(u_\Gamma)$ *be the associated solution to the state system* (9)–(14). *Then the Fréchet derivative* $D\mathcal{S}(u_\Gamma) \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ *is given as follows: if* $h_\Gamma \in \mathcal{X}$, *then* $D\mathcal{S}(u_\Gamma)h_\Gamma = (\xi, \xi_\Gamma)$, *where* (ξ, ξ_Γ, ζ) *with*

$$\xi \in W^{1,\infty}(0,T;H) \cap H^1(0,T;V) \cap L^\infty(0,T;H^2(\Omega)), \tag{52}$$

$$\xi_\Gamma \in W^{1,\infty}(0,T;H_\Gamma) \cap H^1(0,T;V_\Gamma) \cap L^\infty(0,T;H^2(\Gamma)), \tag{53}$$

$$\zeta \in L^\infty(0,T;H^2(\Omega)), \tag{54}$$

is the unique solution to the linearized system

$$\partial_t \xi - \Delta \zeta = 0 \quad \text{in } Q, \tag{55}$$

$$\zeta = \partial_t \xi - \Delta \xi + f''(y)\,\xi \quad \text{in } Q, \tag{56}$$

$$\partial_{\mathbf{n}} \zeta = 0 \quad \text{on } \Sigma, \tag{57}$$

$$\xi_\Gamma = \xi_{|\Gamma} \quad \text{on } \Sigma, \tag{58}$$

$$\partial_t \xi_\Gamma + \partial_{\mathbf{n}} \xi_\Gamma - \Delta_\Gamma \xi_\Gamma + f_\Gamma''(y_\Gamma)\,\xi_\Gamma = h_\Gamma \quad \text{on } \Sigma, \tag{59}$$

$$\xi(0) = 0 \quad \text{in } \Omega, \quad \xi_\Gamma(0) = 0 \quad \text{on } \Gamma. \tag{60}$$

(iii) *The mapping* $D\mathcal{S} : \mathcal{U} \to \mathcal{L}(\mathcal{X}, \mathcal{Y})$, $u_\Gamma \mapsto D\mathcal{S}(u_\Gamma)$, *is Lipschitz continuous on* $\mathcal{U}$ *in the following sense: there is a constant* $K_4^* > 0$, *which depends only on the data and the constant* R, *such that for all* $u_{1,\Gamma}, u_{2,\Gamma} \in \mathcal{U}$ *and all* $h_\Gamma \in \mathcal{X}$ *it holds that*

$$\|(D\mathcal{S}(u_{1,\Gamma}) - D\mathcal{S}(u_{2,\Gamma}))h_\Gamma\|_{\mathcal{Y}} \le K_4^* \, \|u_{1,\Gamma} - u_{2,\Gamma}\|_{L^2(\Sigma)} \, \|h_\Gamma\|_{L^2(\Sigma)}. \tag{61}$$

Proof. First observe that the system (55)–(60) is of form (39)–(44), and with $(\chi, \chi_\Gamma, \mu) := (\xi, \xi_\Gamma, \zeta)$, $g \equiv 0$, $g_\Gamma := h_\Gamma$, and $(\lambda, \lambda_\Gamma) := (f''(y), f_\Gamma''(y_\Gamma))$, the assumptions of Lemma 1 are fulfilled. Consequently, for every $h_\Gamma \in \mathcal{X}$, there is a unique triple (ξ, ξ_Γ, ζ) that satisfies the corresponding variational system (45)–(47) and whose components have the regularity properties in (52), (53) and (54). We may therefore apply [2, Thm. 4.2] to conclude the validity of the assertions (i) and (ii).

It remains to show (iii). To this end, let $u_\Gamma \in \mathcal{U}$ be arbitrary and let $k_\Gamma \in \mathcal{X}$ be such that $u_\Gamma + k_\Gamma \in \mathcal{U}$. We denote $(y^k, y_\Gamma^k) = \mathcal{S}(u_\Gamma + k_\Gamma)$ and

$(y, y_\Gamma) = \mathcal{S}(u_\Gamma)$, and we assume that any $h_\Gamma \in \mathcal{X}$ with $\|h_\Gamma\|_{\mathcal{X}} = 1$ is given. It then suffices to show that there is some $L > 0$, independent of h_Γ, u_Γ and k_Γ, such that

$$\|(\xi^k, \xi_\Gamma^k) - (\xi, \xi_\Gamma)\|_{\mathcal{Y}} \,\le\, L\,\|k_\Gamma\|_{L^2(\Sigma)}\,, \tag{62}$$

where $(\xi^k, \xi_\Gamma^k) = D\mathcal{S}(u_\Gamma + k_\Gamma)h_\Gamma$ and $(\xi, \xi_\Gamma) = D\mathcal{S}(u_\Gamma)h_\Gamma$. For this purpose, in the following we denote by C_i, $i \in \mathbb{N}$, positive constants that depend neither on u_Γ, k_Γ nor on the special choice of $h_\Gamma \in \mathcal{X}$ with $\|h_\Gamma\|_{\mathcal{X}} = 1$. To begin with, observe that the triple $(\widehat{\xi}, \widehat{\xi}_\Gamma, \widehat{\zeta}) := (\xi^k, \xi_\Gamma^k, \zeta^k) - (\xi, \xi_\Gamma, \zeta)$ is the unique solution to the variational analogue of the initial-boundary value problem

$$\partial_t \widehat{\xi} - \Delta \widehat{\zeta} = 0 \quad \text{in } Q, \tag{63}$$

$$\widehat{\zeta} = \partial_t \widehat{\xi} - \Delta \widehat{\xi} + f''(y)\,\widehat{\xi} + \xi^k (f''(y^k) - f''(y)) \quad \text{in } Q, \tag{64}$$

$$\partial_{\mathbf{n}} \widehat{\zeta} = 0 \quad \text{on } \Sigma, \tag{65}$$

$$\widehat{\xi}_\Gamma = \widehat{\xi}_{|\Gamma} \quad \text{on } \Sigma, \tag{66}$$

$$\partial_t \widehat{\xi}_\Gamma + \partial_{\mathbf{n}} \widehat{\xi} - \Delta_\Gamma \widehat{\xi}_\Gamma + f_\Gamma''(y_\Gamma)\,\widehat{\xi}_\Gamma = -\xi_\Gamma^k (f_\Gamma''(y_\Gamma^k) - f_\Gamma''(y_\Gamma)) \quad \text{on } \Sigma, \tag{67}$$

$$\widehat{\xi}(0) = 0 \quad \text{in } \Omega, \quad \widehat{\xi}_\Gamma(0) = 0 \quad \text{on } \Gamma. \tag{68}$$

Moreover, the components of $(\widehat{\xi}, \widehat{\xi}_\Gamma, \widehat{\zeta})$ enjoy the regularity properties indicated in (24), (25), and (28), respectively.

Now observe that it follows from Theorem 1, from part (i) of this proof, and from (38), that $(g, g_\Gamma) := (\xi^k\,(f''(y^k) - f''(y)), -\xi_\Gamma^k\,(f''(y_\Gamma^k) - f''(y_\Gamma)))$ belongs to $H^1(0,T;\mathcal{H}) \cap (L^\infty(Q) \times L^\infty(\Sigma))$, while $(\lambda, \lambda_\Gamma) := (f''(y), f_\Gamma''(y_\Gamma))$ belongs to $W^{1,\infty}(0,T;\mathcal{H}) \cap (L^\infty(Q) \times L^\infty(\Sigma))$. Moreover, (38) also implies that for every $u_\Gamma \in \mathcal{U}$ we have for $(y, y_\Gamma) = \mathcal{S}(u_\Gamma)$ the estimate

$$\|f''(y)\|_{L^\infty(Q)} \,+\, \|f_\Gamma''(y_\Gamma)\|_{L^\infty(\Sigma)} \,\le\, K_1^*\,.$$

Hence, it follows from estimate (48) in Lemma 1 that

$$\begin{aligned}\|(\widehat{\xi}, \widehat{\xi}_\Gamma)\|_{\mathcal{Y}} \,\le\, C_1\big(&\|\xi^k\,(f''(y^k) - f''(y))\|_{L^2(Q)} \\ &+ \|\xi_\Gamma^k\,(f_\Gamma''(y_\Gamma^k) - f_\Gamma''(y_\Gamma))\|_{L^2(\Sigma)}\big).\end{aligned} \tag{69}$$

Now, by the mean value theorem and (38), there exists a positive constant C_2 such that almost everywhere in Q (on Σ, respectively)

$$|f''(y^k) - f''(y)| \,\le\, C_2\,|y^k - y| \quad \text{and} \quad |f_\Gamma''(y_\Gamma^k) - f_\Gamma''(y_\Gamma)| \,\le\, C_2\,|y_\Gamma^k - y_\Gamma|\,. \tag{70}$$

At this point, we recall that $\mathcal{U}$ is a bounded subset of $\mathcal{X}$. Since $u_\Gamma + k_\Gamma \in \mathcal{U}$ and $\|h_\Gamma\|_{\mathcal{X}} = 1$, we thus can infer from (38) and from the estimate (48) in Lemma 1 that (ξ^k, ξ^k_Γ) is bounded in $\mathcal{Y}$ independently of k_Γ, u_Γ and the choice of $h_\Gamma \in \mathcal{X}$ with $\|h_\Gamma\|_{\mathcal{X}} = 1$. Using the embedding $V \subset L^4(\Omega)$ and the stability estimate proved in Theorem 1, we therefore have that

$$\begin{aligned}
&\|\xi^k \, (f''(y^k) - f''(y))\|^2_{L^2(Q)} \,\leq\, C_2 \int_0^T \!\!\int_\Omega \Big(|\xi^k|^2 \, |y^k - y|^2 \Big) \, dx \, dt \\
&\leq\, C_2 \int_0^T \Big(\|\xi^k(t)\|^2_{L^4(\Omega)} \, \|y^k(t) - y(t)\|^2_{L^4(\Omega)} \Big) \, dt \\
&\leq\, C_3 \, \|(y^k, y^k_\Gamma) - (y, y_\Gamma)\|^2_{\mathcal{Y}} \,\leq\, C_4 \, \|k_\Gamma\|^2_{L^2(\Sigma)} \,. \qquad (71)
\end{aligned}$$

Since an analogous estimate holds for the second summand in the bracket on the right-hand side of (69), the assertion follows. □

With the Lipschitz estimate (61) at hand, we are now in the position to show the existence of the second-order Fréchet derivative. We have the following result.

Theorem 3. *Assume that* **(A1)**–**(A4)** *are fulfilled. Then the following holds true:*
(i) *The control-to-state operator $\mathcal{S}$ is twice Fréchet differentiable in $\mathcal{U}$ as a mapping from $\mathcal{U} \subset \mathcal{X}$ to $\mathcal{Y}$.*
(ii) *For all $u_\Gamma \in \mathcal{U}$, the second Fréchet derivative $D^2\mathcal{S}(u_\Gamma) \in \mathcal{L}(\mathcal{X}, \mathcal{L}(\mathcal{X}, \mathcal{Y}))$ is defined as follows: if $h_\Gamma, k_\Gamma \in \mathcal{X}$ are arbitrary, then $D^2\mathcal{S}(u_\Gamma)[h_\Gamma, k_\Gamma] =: (\eta, \eta_\Gamma)$ is the unique solution to the initial-boundary value problem*

$$\partial_t \eta - \Delta \vartheta = 0 \quad \text{in } Q, \qquad (72)$$

$$\vartheta = \partial_t \eta - \Delta \eta + f''(y)\, \eta + f^{(3)}(y)\, \varphi\, \psi \quad \text{in } Q, \qquad (73)$$

$$\partial_{\mathbf{n}} \vartheta = 0 \quad \text{on } \Sigma, \qquad (74)$$

$$\eta_\Gamma = \eta_{|\Gamma} \quad \text{on } \Sigma, \qquad (75)$$

$$\partial_t \eta_\Gamma + \partial_{\mathbf{n}} \eta - \Delta_\Gamma \eta_\Gamma + f''_\Gamma(y_\Gamma)\, \eta_\Gamma = -f^{(3)}_\Gamma(y_\Gamma)\, \varphi_\Gamma\, \psi_\Gamma \quad \text{on } \Sigma, \qquad (76)$$

$$\eta(0) = 0 \quad \text{in } \Omega, \quad \eta_\Gamma(0) = 0 \quad \text{on } \Gamma, \qquad (77)$$

where we have put

$$(y, y_\Gamma) = \mathcal{S}(u_\Gamma), \quad (\varphi, \varphi_\Gamma) = D\mathcal{S}(u_\Gamma) h_\Gamma, \quad (\psi, \psi_\Gamma) = D\mathcal{S}(u_\Gamma) k_\Gamma \,. \qquad (78)$$

(iii) *The mapping* $D^2\mathcal{S} : \mathcal{U} \to \mathcal{L}(\mathcal{X}, \mathcal{L}(\mathcal{X}, \mathcal{Y}))$, $u_\Gamma \mapsto D^2\mathcal{S}(u_\Gamma)$, *is Lipschitz continuous on* $\mathcal{U}$ *in the following sense: there exists a constant* $K_5^* > 0$, *which depends only on the data and on the constant* R, *such that for every* $u_{1,\Gamma}, u_{2,\Gamma} \in \mathcal{U}$ *and all* $h_\Gamma, k_\Gamma \in \mathcal{X}$ *it holds that*

$$\begin{aligned} &\|(D^2\mathcal{S}(u_{1,\Gamma}) - D^2\mathcal{S}(u_{2,\Gamma}))[h_\Gamma, k_\Gamma]\|_{\mathcal{Y}} \\ &\leq K_5^* \, \|u_{1,\Gamma} - u_{2,\Gamma}\|_{L^2(\Sigma)} \, \|h_\Gamma\|_{L^2(\Sigma)} \, \|k_\Gamma\|_{L^2(\Sigma)}. \end{aligned} \tag{79}$$

Proof. At first, it is easily verified that the pair $(g, g_\Gamma) := (f^{(3)}(y)\,\varphi\,\psi, \; -f_\Gamma^{(3)}(y_\Gamma)\,\varphi_\Gamma\,\psi_\Gamma)$ belongs to $H^1(0,T;\mathcal{H}) \cap (L^\infty(Q) \times L^\infty(\Sigma))$. We thus can argue as in the proof of Theorem 2 to deduce from Lemma 1 that the system (72)–(77) is uniquely solvable in the sense that its variational counterpart has a unique solution $(\eta, \eta_\Gamma, \vartheta)$ whose components enjoy the regularity indicated in (24), (25), and (28), respectively. Moreover, by (48) we have the estimate

$$\|(\eta, \eta_\Gamma)\|_{\mathcal{Y}} \leq C_1 \left(\left\| f^{(3)}(y)\,\varphi\,\psi \right\|_{L^2(Q)} + \left\| f_\Gamma^{(3)}(y_\Gamma)\,\varphi_\Gamma\,\psi_\Gamma \right\|_{L^2(\Sigma)} \right). \tag{80}$$

Here, and in the remainder of the proof of parts (i), (ii), we denote by C_i, $i \in \mathbb{N}$, positive constants that do not depend on the quantities h_Γ, k_Γ, and u_Γ. Using (38), and invoking the embedding $V \subset L^4(\Omega)$, we find that

$$\begin{aligned} &\left\| f^{(3)}(y)\,\varphi\,\psi \right\|_{L^2(Q)}^2 \leq C_2 \int_0^T \!\! \int_\Omega |\varphi|^2 \, |\psi|^2 \, dx \, dt \\ &\leq C_2 \int_0^T \|\varphi(t)\|_{L^4(\Omega)}^2 \, \|\psi(t)\|_{L^4(\Omega)}^2 \, dt \leq C_3 \, \|\varphi\|_{L^\infty(0,T;V)}^2 \, \|\psi\|_{L^\infty(0,T;V)}^2 \\ &\leq C_4 \, \|h_\Gamma\|_{L^2(\Sigma)}^2 \, \|k_\Gamma\|_{L^2(\Sigma)}^2 , \end{aligned} \tag{81}$$

where the validity of the last inequality can be seen as follows: by definition (recall (78)), $(\varphi, \varphi_\Gamma)$ is the unique solution to the linear problem (55)–(60). We can therefore infer from (48) that $\|(\varphi, \varphi_\Gamma)\|_{\mathcal{Y}} \leq C_5 \, \|h_\Gamma\|_{L^2(\Sigma)}$. By the same token, we conclude that $\|(\psi, \psi_\Gamma)\|_{\mathcal{Y}} \leq C_6 \, \|k_\Gamma\|_{L^2(\Sigma)}$. The asserted inequality therefore follows from the definition of the norm of the space $\mathcal{Y}$, and we obtain from similar reasoning that also

$$\left\| f_\Gamma^{(3)}(y_\Gamma)\,\varphi_\Gamma\,\psi_\Gamma \right\|_{L^2(\Sigma)} \leq C_7 \, \|h_\Gamma\|_{L^2(\Sigma)} \, \|k_\Gamma\|_{L^2(\Sigma)} .$$

Hence, we get

$$\|(\eta, \eta_\Gamma)\|_{\mathcal{Y}} \leq C_8 \, \|h_\Gamma\|_{L^2(\Sigma)} \, \|k_\Gamma\|_{L^2(\Sigma)} . \tag{82}$$

In particular, it follows that the bilinear mapping $\mathcal{X} \times \mathcal{X} \mapsto \mathcal{Y}$, $[k_\Gamma, h_\Gamma] \mapsto (\eta, \eta_\Gamma)$, is continuous.

Now we prove the assertions concerning existence and form of the second Fréchet derivative. Since $\mathcal{U}$ is open, there is some $\Lambda > 0$ such that $u_\Gamma + k_\Gamma \in \mathcal{U}$ whenever $\|k_\Gamma\|_{\mathcal{X}} \leq \Lambda$. In the following, we only consider such perturbations $k_\Gamma \in \mathcal{X}$. We observe that for $(y, y_\Gamma) = \mathcal{S}(u_\Gamma)$ and for $(y^k, y_\Gamma^k) = \mathcal{S}(u_\Gamma + k_\Gamma)$ the global estimates (34)–(36) and (38) are satisfied.

After these preparations, we notice that it suffices to show that

$$\begin{aligned} &\big\| D\mathcal{S}(u_\Gamma + k_\Gamma) - D\mathcal{S}(u_\Gamma) - D^2\mathcal{S}(u_\Gamma) k_\Gamma \big\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} \\ &= \sup_{\|h_\Gamma\|_{\mathcal{X}}=1} \big\| \big(D\mathcal{S}(u_\Gamma + k_\Gamma) - D\mathcal{S}(u_\Gamma) - D^2\mathcal{S}(u_\Gamma) k_\Gamma \big)\, h_\Gamma \big\|_{\mathcal{Y}} \\ &\leq \overline{C}\, \|k_\Gamma\|^2_{L^2(\Sigma)} \end{aligned} \tag{83}$$

with a constant $\overline{C}$ independent of k_Γ.

To this end, let $h_\Gamma \in \mathcal{X}$ be arbitrary with $\|h_\Gamma\|_{\mathcal{X}} = 1$. We put $(\rho, \rho_\Gamma) = D\mathcal{S}(u_\Gamma + k_\Gamma) h_\Gamma$, define the pairs $(\varphi, \varphi_\Gamma), (\psi, \psi_\Gamma)$ as in (78), and define

$$(\nu, \nu_\Gamma) := (\rho, \rho_\Gamma) - (\varphi, \varphi_\Gamma) - (\eta, \eta_\Gamma).$$

Observe that the components of (ν, ν_Γ) have the regularity properties indicated in (24) and (25), respectively. Moreover, in view of (83), we need to show that

$$\|(\nu, \nu_\Gamma)\|_{\mathcal{Y}} \leq \overline{C}\, \|k_\Gamma\|^2_{L^2(\Sigma)}\,. \tag{84}$$

Now, invoking the explicit expressions for the quantities defined above, it is easily seen that the triple (ν, ν_Γ, π) (where π is defined below) is the unique solution to the variational counterpart of the linear initial-boundary value problem

$$\partial_t \nu - \Delta \pi = 0 \quad \text{in } Q, \tag{85}$$

$$\pi = \partial_t \nu - \Delta \nu + f''(y)\, \nu + \sigma \quad \text{in } Q, \tag{86}$$

$$\partial_{\mathbf{n}} \pi = 0 \quad \text{on } \Sigma, \tag{87}$$

$$\nu_\Gamma = \nu_{|\Gamma} \quad \text{and} \quad \partial_t \nu_\Gamma + \partial_{\mathbf{n}} \nu - \Delta_\Gamma \nu_\Gamma + f_\Gamma''(y_\Gamma)\, \nu_\Gamma = \sigma_\Gamma \quad \text{on } \Sigma, \tag{88}$$

$$\nu(0) = 0 \quad \text{in } \Omega, \quad \nu_\Gamma(0) = 0 \quad \text{on } \Gamma, \tag{89}$$

where we have put

$$\sigma := \rho \left(f''(y^k) - f''(y) \right) \, - \, f^{(3)}(y) \, \varphi \, \psi,$$
$$\sigma_\Gamma := -\rho_\Gamma \left(f''_\Gamma(y^k_\Gamma) - f''_\Gamma(y_\Gamma) \right) \, + \, f^{(3)}_\Gamma(y_\Gamma) \, \varphi_\Gamma \, \psi_\Gamma \, . \tag{90}$$

In view of (38), and since it is easily checked that (σ, σ_Γ) belongs to the space $H^1(0,T;\mathcal{H}) \cap (L^\infty(Q) \times L^\infty(\Sigma))$, we may again invoke the estimate (48) in Lemma 1 to conclude that (84) is satisfied if only

$$\|(\sigma, \sigma_\Gamma)\|_{L^2(0,T;\mathcal{H})} \, \leq \, \overline{C} \, \|k_\Gamma\|^2_{L^2(\Sigma)} \, . \tag{91}$$

Applying Taylor's theorem to f'', and recalling (38), we readily see that there is a function $\omega_f \in L^\infty(Q)$ such that, a. e. in Q,

$$f''(y^k) - f''(y) = f^{(3)}(y) \, (y^k - y - \psi) \, + \, f^{(3)}(y) \, \psi \, + \, \omega_f \, (y^k - y)^2 \, . \tag{92}$$

Hence, we have that

$$\sigma \, = \rho \, f^{(3)}(y) \, (y^k - y - \psi) \, + \, \psi \, f^{(3)}(y) \, (\rho - \varphi) \, + \, \rho \, \omega_f \, (y^k - y)^2 \, . \tag{93}$$

Now observe that from the proof of Fréchet differentiability (see inequality (4.5) in the proof of [2, Thm. 4.2]) and from (61) we can conclude the estimates

$$\|(y^k, y^k_\Gamma) - (y, y_\Gamma) - (\psi, \psi_\Gamma)\|_{\mathcal{Y}} \, \leq \, C_9 \, \|k_\Gamma\|^2_{L^2(\Sigma)} \, ,$$
$$\|(\rho, \rho_\Gamma) - (\varphi, \varphi_\Gamma)\|_{\mathcal{Y}} \, \leq \, C_{10} \, \|k_\Gamma\|_{L^2(\Sigma)} \, . \tag{94}$$

Moreover, we can infer from inequality (36) in Theorem 1 that

$$\|(y^k, y^k_\Gamma) - (y, y_\Gamma)\|_{\mathcal{Y}} \, \leq \, K_2^* \, \|k_\Gamma\|_{L^2(\Sigma)} \, , \tag{95}$$

and it follows from Lemma 1 that (ρ, ρ_Γ) is bounded in $\mathcal{Y}$ by a positive constant that is independent of $k_\Gamma, h_\Gamma \in \mathcal{X}$ with $\|k_\Gamma\|_{\mathcal{X}} \leq \Lambda$ and $\|h_\Gamma\|_{\mathcal{X}} = 1$. Finally, we conclude from Lemma 1 (ii) that with a suitable constant $C_{11} > 0$ it holds

$$\|(\psi, \psi_\Gamma)\|_{\mathcal{Y}} \, \leq \, C_{11} \, \|k_\Gamma\|_{L^2(\Sigma)} \, . \tag{96}$$

After these preparations, and invoking Hölder's inequality and the continuity of the embeddings $V \subset L^4(\Omega)$ and $V \subset L^6(\Omega)$, we can estimate as follows:

$$\|\sigma\|^2_{L^2(Q)} \, \leq \, C_{12} \int_0^T \int_\Omega \big(|\rho|^2 \, |y^k - y - \psi|^2 + |\psi|^2 |\rho - \varphi|^2 + |\rho|^2 |y^k - y|^4 \big) dx \, dt$$

$$
\begin{aligned}
&\le C_{12} \int_0^T \Big(\|\rho(t)\|^2_{L^4(\Omega)} \, \|(y^k - y - \psi)(t)\|^2_{L^4(\Omega)} \\
&\quad + \|\psi(t)\|^2_{L^4(\Omega)} \, \|\rho(t) - \varphi(t)\|^2_{L^4(\Omega)} \, + \, \|\rho(t)\|^2_{L^6(\Omega)} \, \|y^k(t) - y(t)\|^4_{L^6(\Omega)} \Big) \, dt \\
&\le C_{13} \sup_{t \in (0,T)} \Big(\|\rho(t)\|^2_V \, \|(y^k - y - \psi)(t)\|^2_V \, + \, \|\psi(t)\|^2_V \, \|\rho(t) - \varphi(t)\|^2_V \\
&\qquad\qquad + \|\rho(t)\|^2_V \, \|y^k(t) - y(t)\|^4_V \Big) \\
&\le C_{14} \, \|k_\Gamma\|^4_{L^2(\Sigma)}. \qquad (97)
\end{aligned}
$$

By the same reasoning, a similar estimate can be derived for $\|\sigma_\Gamma\|_{L^2(\Sigma)}$, which concludes the proof of the assertions (i) and (ii).

Next, we prove the assertion (iii). To this end, suppose that $u_\Gamma \in \mathcal{U}$ and that h_Γ and k_Γ are arbitrarily chosen in $\mathcal{X}$, and let $\delta_\Gamma \in \mathcal{X}$ be arbitrary with $u_\Gamma + \delta_\Gamma \in \mathcal{U}$. In the following, we will denote by C_i, $i \in \mathbb{N}$, positive constants that do not depend on any of these quantities. We put

$$
\begin{aligned}
&(y, y_\Gamma) = \mathcal{S}(u_\Gamma), \quad (y^\delta, y^\delta_\Gamma) = \mathcal{S}(u_\Gamma + \delta_\Gamma), \\
&(\varphi, \varphi_\Gamma) = D\mathcal{S}(u_\Gamma) h_\Gamma, \quad (\varphi^\delta, \varphi^\delta_\Gamma) = D\mathcal{S}(u_\Gamma + \delta_\Gamma) h_\Gamma, \\
&(\psi, \psi_\Gamma) = D\mathcal{S}(u_\Gamma) k_\Gamma, \quad (\psi^\delta, \psi^\delta_\Gamma) = D\mathcal{S}(u_\Gamma + \delta_\Gamma) k_\Gamma, \\
&(\eta, \eta_\Gamma) = D^2\mathcal{S}(u_\Gamma)[h_\Gamma, k_\Gamma], \quad (\eta^\delta, \eta^\delta_\Gamma) = D^2\mathcal{S}(u_\Gamma + \delta_\Gamma)[h_\Gamma, k_\Gamma] \, .
\end{aligned}
$$

From the previous results, in particular, (36) and (61), we can infer that there is a constant $C_1 > 0$ such that

$$
\begin{aligned}
&\|(\varphi, \varphi_\Gamma)\|_{\mathcal{Y}} + \|(\varphi^\delta, \varphi^\delta_\Gamma)\|_{\mathcal{Y}} \, \le \, C_1 \, \|h_\Gamma\|_{L^2(\Sigma)}, \\
&\|(\psi, \psi_\Gamma)\|_{\mathcal{Y}} + \|(\psi^\delta, \psi^\delta_\Gamma)\|_{\mathcal{Y}} \, \le \, C_1 \, \|k_\Gamma\|_{L^2(\Sigma)}, \\
&\|(\eta, \eta_\Gamma)\|_{\mathcal{Y}} \, + \, \|(\eta^\delta, \eta^\delta_\Gamma)\|_{\mathcal{Y}} \, \le \, C_1 \, \|h_\Gamma\|_{L^2(\Sigma)} \, \|k_\Gamma\|_{L^2(\Sigma)}, \\
&\|(y^\delta, y^\delta_\Gamma) - (y, y_\Gamma)\|_{\mathcal{Y}} \, \le \, C_1 \, \|\delta_\Gamma\|_{L^2(\Sigma)}, \\
&\|(\varphi^\delta, \varphi^\delta_\Gamma) - (\varphi, \varphi_\Gamma)\|_{\mathcal{Y}} \, \le \, C_1 \, \|\delta_\Gamma\|_{L^2(\Sigma)} \, \|h_\Gamma\|_{L^2(\Sigma)}, \\
&\|(\psi^\delta, \psi^\delta_\Gamma) - (\psi, \psi_\Gamma)\|_{\mathcal{Y}} \, \le \, C_1 \, \|\delta_\Gamma\|_{L^2(\Sigma)} \, \|k_\Gamma\|_{L^2(\Sigma)} \, . \qquad (98)
\end{aligned}
$$

Now observe that $(\tilde\eta, \tilde\eta_\Gamma) = (\eta^\delta, \eta^\delta_\Gamma) - (\eta, \eta_\Gamma)$ and $\tilde\vartheta = \vartheta^\delta - \vartheta$ (where ϑ^δ and ϑ have their obvious meaning corresponding to (73)) satisfy the linear

initial-boundary value problem

$$\partial_t \tilde\eta - \Delta \tilde\vartheta = 0 \quad \text{in } Q, \tag{99}$$

$$\tilde\vartheta = \partial_t \tilde\eta - \Delta \tilde\eta + f''(y)\,\tilde\eta + \sigma \quad \text{in } Q, \tag{100}$$

$$\partial_{\mathbf{n}} \tilde\vartheta = 0 \quad \text{on } \Sigma, \tag{101}$$

$$\tilde\eta_\Gamma = \tilde\eta_{|\Gamma} \quad \text{and} \quad \partial_t \tilde\eta_\Gamma + \partial_{\mathbf{n}} \tilde\eta - \Delta_\Gamma \tilde\eta_\Gamma + f_\Gamma''(y_\Gamma)\,\tilde\eta_\Gamma = \sigma_\Gamma \quad \text{on } \Sigma, \tag{102}$$

$$\tilde\eta(0) = 0 \quad \text{in } \Omega, \quad \tilde\eta_\Gamma(0) = 0 \quad \text{on } \Gamma, \tag{103}$$

where we have put

$$\begin{aligned} \sigma &= \eta^\delta (f''(y^\delta) - f''(y)) + (f^{(3)}(y^\delta)\,\varphi^\delta\,\psi^\delta - f^{(3)}(y)\,\varphi\,\psi)\,, \\ \sigma_\Gamma &= -\eta_\Gamma^\delta (f_\Gamma''(y_\Gamma^\delta) - f_\Gamma''(y_\Gamma)) - (f_\Gamma^{(3)}(y_\Gamma^\delta)\,\varphi_\Gamma^\delta\,\psi_\Gamma^\delta - f_\Gamma^{(3)}(y_\Gamma)\,\varphi_\Gamma\,\psi_\Gamma)\,. \end{aligned} \tag{104}$$

The system (99)–(103) is again of the form (39)–(44), and since it is readily verified that (σ, σ_Γ) belongs to the space $H^1(0,T;\mathcal{H}) \cap (L^\infty(Q) \times L^\infty(\Sigma))$, we may employ Lemma 1 once more to conclude that

$$\|(\tilde\eta, \tilde\eta_\Gamma)\|_{\mathcal{Y}} \,\le\, C_2\, \|(\sigma, \sigma_\Gamma)\|_{L^2(0,T;\mathcal{H})}\,, \tag{105}$$

so that it remains to show an estimate of the form

$$\|(\sigma, \sigma_\Gamma)\|_{L^2(0,T;\mathcal{H})} \,\le\, C_3\, \|\delta_\Gamma\|_{L^2(\Sigma)}\, \|h_\Gamma\|_{L^2(\Sigma)}\, \|k_\Gamma\|_{L^2(\Sigma)}\,. \tag{106}$$

Since

$$\begin{aligned} &f^{(3)}(y^\delta)\,\varphi^\delta\,\psi^\delta - f^{(3)}(y)\,\varphi\,\psi \\ &= \varphi^\delta\,\psi\,(f^{(3)}(y^\delta) - f^{(3)}(y)) + f^{(3)}(y^\delta)\,\varphi^\delta\,(\psi^\delta - \psi) + f^{(3)}(y)\,\psi\,(\varphi^\delta - \varphi)\,, \end{aligned} \tag{107}$$

we can infer from (38) that, almost everywhere in Q,

$$|\sigma| \,\le\, C_4\,(|\eta^\delta|\,|y^\delta - y| + |\varphi^\delta|\,|\psi|\,|y^\delta - y| + |\varphi^\delta|\,|\psi^\delta - \psi| + |\psi|\,|\varphi^\delta - \varphi|)\,. \tag{108}$$

Using (98), Hölder's inequality and the continuity of the embedding $V \subset L^4(\Omega)$, we find that

$$\begin{aligned} \int_0^T \!\!\int_\Omega \Big(|\eta^\delta|^2\,|y^\delta - y|^2\Big)\,dx\,dt \,&\le\, \int_0^T \Big(\|\eta^\delta(t)\|_{L^4(\Omega)}^2\,\|(y^\delta - y)(t)\|_{L^4(\Omega)}^2\Big)\,dt \\ &\le\, C_5\,\|\eta^\delta\|_{L^\infty(0,T;V)}^2 \|y^\delta - y\|_{L^\infty(0,T;V)}^2 \\ &\le\, C_6\,\|\delta_\Gamma\|_{L^2(\Sigma)}^2\,\|h_\Gamma\|_{L^2(\Sigma)}^2\,\|k_\Gamma\|_{L^2(\Sigma)}^2\,. \end{aligned} \tag{109}$$

Similar reasoning yields

$$\begin{aligned} &\|\varphi^\delta(\psi^\delta - \psi)\|^2_{L^2(Q)} + \|\psi(\varphi^\delta - \varphi)\|^2_{L^2(Q)} \\ &\quad \leq C_7 \, \|\delta_\Gamma\|^2_{L^2(\Sigma)} \, \|h_\Gamma\|^2_{L^2(\Sigma)} \, \|k_\Gamma\|^2_{L^2(\Sigma)} \,. \end{aligned} \tag{110}$$

Moreover, once again invoking (98), Hölder's inequality, and the continuity of the embedding $V \subset L^6(\Omega)$, we conclude that

$$\begin{aligned} &\int_0^T \!\! \int_\Omega \Big(|\varphi^\delta|^2 \, |\psi|^2 \, |y^\delta - y|^2 \Big) \, dx \, dt \\ &\leq \int_0^T \Big(\|(y^\delta - y)(t)\|^2_{L^6(\Omega)} \, \|\varphi^\delta(t)\|^2_{L^6(\Omega)} \, \|\psi(t)\|^2_{L^6(\Omega)} \Big) \, dt \\ &\leq C_8 \, \|\varphi^\delta\|^2_{L^\infty(0,T;V)} \, \|\psi\|^2_{L^\infty(0,T;V)} \, \|y^\delta - y\|^2_{L^\infty(0,T;V)} \\ &\leq C_9 \, \|\delta_\Gamma\|^2_{L^2(\Sigma)} \, \|h_\Gamma\|^2_{L^2(\Sigma)} \, \|k_\Gamma\|^2_{L^2(\Sigma)} \,. \end{aligned} \tag{111}$$

Finally, we can estimate $\|\sigma_\Gamma\|_{L^2(\Sigma)}$, deriving estimates similar to (108)–(111), which entails the validity of the required estimate (106). With this, the assertion is completely proved. □

4 Optimality conditions

Now that the second-order Fréchet derivative of the control-to-state operator for problem **(CP)** is obtained, we can address the matter of deriving second-order sufficient optimality conditions. As a preparation of the corresponding theorem, we provide the adjoint system and the first-order necessary optimality conditions. Since these were already established in [2], we only present the results without proofs.

At first, it is easily shown (cf. [2, Thm. 2.2]) that **(CP)** has a solution. For the remainder of this paper, let us assume that $\bar{u}_\Gamma \in \mathcal{U}_{ad}$ is any such minimizer and that $(\bar{y}, \bar{y}_\Gamma, \bar{w})$, where $(\bar{y}, \bar{y}_\Gamma) = \mathcal{S}(\bar{u}_\Gamma)$, is the associated solution to the state system. Recall that $(\bar{y}, \bar{y}_\Gamma, \bar{w})$ has the regularity properties (24), (25), and (28), respectively, and that (38) is satisfied for $(y, y_\Gamma) = (\bar{y}, \bar{y}_\Gamma)$.

The adjoint system to the problem **(CP)** is formally given by

$$q + \Delta p = 0 \quad \text{in } Q, \tag{112}$$

$$-\partial_t(p+q) - \Delta q + f''(\bar{y})\, q = b_Q(\bar{y} - z_Q) \quad \text{in } Q, \tag{113}$$

$$\partial_{\mathbf{n}} p = 0 \quad \text{on } \Sigma, \tag{114}$$

$$q_\Gamma = q_{|\Gamma} \quad \text{and} \quad -\partial_t q_\Gamma + \partial_{\mathbf{n}} q - \Delta_\Gamma q_\Gamma + f_\Gamma''(\bar{y}_\Gamma)\, q_\Gamma = b_\Sigma(\bar{y}_\Gamma - z_\Sigma) \quad \text{on } \Sigma, \tag{115}$$

$$(p+q)(T) = b_\Omega(\bar{y}(T) - z_\Omega) \quad \text{in } \Omega, \tag{116}$$

$$q_\Gamma(T) = b_\Gamma(\bar{y}_\Gamma(T) - z_\Gamma) \quad \text{on } \Gamma, \tag{117}$$

and was derived in [2] under the additional compatibility assumption

$$b_\Omega = b_\Gamma = 0. \tag{118}$$

In order to keep the technicalities at a reasonable level, we will from now on always assume that (118) is fulfilled; we remark that in [2, Remark 5.6] it has been pointed out that this assumption is dispensable at the expense of less regularity of the adjoint state variables.

The following result was proved in [2, Thm. 2.4].

Theorem 4. *Let* **(A1)**–**(A4)** *and* (118) *be fulfilled. Then the adjoint system* (112)–(117) *has a unique solution in the following sense: there is a unique triple* (p, q, q_Γ) *with the regularity properties*

$$p \in H^1(0,T; H^2(\Omega)) \cap L^2(0,T; H^4(\Omega)), \tag{119}$$

$$q \in H^1(0,T; H) \cap L^2(0,T; H^2(\Omega)), \tag{120}$$

$$q_\Gamma \in H^1(0,T; H_\Gamma) \cap L^2(0,T; H^2(\Gamma)), \tag{121}$$

$$q_\Gamma(t) = q(t)_{|\Gamma} \quad \textit{for a.a. } t \in (0,T), \tag{122}$$

that solves for a.a. $t \in (0,T)$ *the variational equations*

$$\int_\Omega q(t)\, v\, dx = \int_\Omega \nabla p(t) \cdot \nabla v\, dx \quad \forall\, v \in V, \tag{123}$$

$$\begin{aligned} &-\int_\Omega \partial_t\big(p(t)+q(t)\big)\, v\, dx + \int_\Omega \nabla q(t) \cdot \nabla v\, dx + \int_\Omega f''(\bar{y}(t))\, q(t)\, v\, dx \\ &\quad - \int_\Gamma \partial_t q_\Gamma(t)\, v_\Gamma\, d\Gamma + \int_\Gamma \nabla_\Gamma q_\Gamma(t) \cdot \nabla_\Gamma v_\Gamma\, d\Gamma + \int_\Gamma f_\Gamma''(\bar{y}_\Gamma(t))\, q_\Gamma(t)\, v_\Gamma\, d\Gamma \\ &= \int_\Omega b_Q\big(\bar{y}(t) - z_Q(t)\big) v\, dx + \int_\Gamma b_\Sigma\big(\bar{y}_\Gamma(t) - z_\Sigma(t)\big) v_\Gamma\, d\Gamma \\ &\quad \textit{for all } (v, v_\Gamma) \in \mathcal{V}, \end{aligned} \tag{124}$$

and the final condition

$$\int_\Omega (p+q)(T)\, v\, dx + \int_\Gamma q_\Gamma(T)\, v_\Gamma\, d\Gamma = 0 \quad \forall (v, v_\Gamma) \in \mathcal{V}\,. \tag{125}$$

Now, let us introduce the "reduced cost functional" $\widetilde{\mathcal{J}} : \mathcal{U} \to \mathbb{R}$ by

$$\widetilde{\mathcal{J}}(u_\Gamma) := \mathcal{J}(y, y_\Gamma, u_\Gamma), \quad \text{where} \quad (y, y_\Gamma) = \mathcal{S}(u_\Gamma). \tag{126}$$

Since $\bar{u}_\Gamma$ is an optimal control with associated optimal state $(\bar{y}, \bar{y}_\Gamma) = \mathcal{S}(\bar{u}_\Gamma)$, the necessary condition for optimality is

$$D\widetilde{\mathcal{J}}(\bar{u}_\Gamma)(v_\Gamma - \bar{u}_\Gamma) \,\geq 0 \quad \text{for every } v_\Gamma \in \mathcal{U}_{ad}, \tag{127}$$

or, written explicitly (recall that $b_\Omega = b_\Gamma = 0$),

$$\begin{aligned} & b_Q \int_0^T \!\! \int_\Omega (\bar{y} - z_Q)\, \xi \, dx \, dt \,+\, b_\Sigma \int_0^T \!\! \int_\Gamma (\bar{y}_\Gamma - z_\Sigma)\, \xi_\Gamma \, d\Gamma \, dt \\ & +\, b_0 \int_0^T \!\! \int_\Gamma \bar{u}_\Gamma \, (v_\Gamma - \bar{u}_\Gamma) \, d\Gamma \, dt \,\geq\, 0 \quad \text{for every } \, v_\Gamma \in \mathcal{U}_{ad}, \end{aligned} \tag{128}$$

where, for any given $v_\Gamma \in \mathcal{U}_{ad}$, the functions ξ, ξ_Γ are the first two components of the solution triple (ξ, ξ_Γ, ζ) to the linearized problem (55)–(60) associated with $h_\Gamma = v_\Gamma - \bar{u}_\Gamma$. Moreover, since the adjoint variables have been constructed in such a way that

$$b_Q \int_0^T \!\! \int_\Omega (\bar{y} - z_Q) \xi \, dx \, dt + b_\Sigma \int_0^T \!\! \int_\Gamma (\bar{y}_\Gamma - z_\Sigma) \xi_\Gamma \, d\Gamma \, dt = \int_0^T \!\! \int_\Gamma q_\Gamma (v_\Gamma - \bar{u}_\Gamma) d\Gamma \, dt, \tag{129}$$

we can rewrite (128) in the form (see also [2, Thm. 2.5])

$$\int_0^T \!\! \int_\Gamma (q_\Gamma + b_0 \, \bar{u}_\Gamma)(v_\Gamma - \bar{u}_\Gamma) \, d\Gamma \, dt \,\geq\, 0 \quad \text{for every } \, v_\Gamma \in \mathcal{U}_{ad}. \tag{130}$$

In particular, if $b_0 > 0$, then $\bar{u}_\Gamma$ is the orthogonal projection of $-q_\Gamma / b_0$ onto $\mathcal{U}_{ad}$ with respect to the standard scalar product in $L^2(\Sigma)$.

After these preparations, we now derive sufficient conditions for optimality. But, since the control-to-state operator $\mathcal{S}$ is not Fréchet differentiable on $L^2(\Sigma)$ but only on $\mathcal{U} \subset \mathcal{X}$, we are faced with the so-called "two-norm discrepancy", which makes it impossible to establish second-order sufficient optimality conditions by means of the same simple arguments as in the finite-dimensional case or, e. g., in the proof of [6, Thm. 4.23, p. 231]. It will thus be necessary to tailor the conditions in such a way as to overcome the two-norm discrepancy. At the same time, for practical purposes the conditions should not be overly restrictive. For such an approach, we follow the lines of Chapter 5 in [6], here. Since many of the arguments developed here

are rather similar to those employed in [6], we can afford to be sketchy and refer the reader to [6] for full details.

To begin with, the quadratic cost functional $\mathcal{J}$, viewed as a map from $C^0([0,T];\mathcal{H}) \times \mathcal{U}$ into $\mathbb{R}$, is obviously twice continuously Fréchet differentiable on $C^0([0,T];\mathcal{H}) \times \mathcal{U}$ and thus, in particular, at $((\bar y, \bar y_\Gamma), \bar u_\Gamma)$. Moreover, since $b_\Omega = b_\Gamma = 0$, we have for any $((y, y_\Gamma), u_\Gamma) \in C^0([0,T];\mathcal{H}) \times \mathcal{U}$ and any $((v, v_\Gamma), h_\Gamma), ((\omega, \omega_\Gamma), k_\Gamma) \in C^0([0,T];\mathcal{H}) \times \mathcal{X}$ that

$$\begin{aligned} &D^2\mathcal{J}((y, y_\Gamma), u_\Gamma)[((v, v_\Gamma), h_\Gamma), ((\omega, \omega_\Gamma), k_\Gamma)] \\ &= b_Q \int_0^T \int_\Omega v\,\omega \,dx\,dt \,+\, b_\Sigma \int_0^T \int_\Gamma v_\Gamma\,\omega_\Gamma \,d\Gamma\,dt \,+\, b_0 \int_0^T \int_\Gamma h_\Gamma\,k_\Gamma \,d\Gamma\,dt\,. \end{aligned} \tag{131}$$

It then follows from Theorem 3 and from the chain rule that the reduced cost functional $\tilde{\mathcal{J}}$ is also twice continuously Fréchet differentiable on $\mathcal{U}$. Now let $h_\Gamma, k_\Gamma \in \mathcal{X}$ be arbitrary. In accordance with our previous notation, we put

$$(\varphi, \varphi_\Gamma) = D\mathcal{S}(\bar u_\Gamma)h_\Gamma, \;\; (\psi, \psi_\Gamma) = D\mathcal{S}(\bar u_\Gamma)k_\Gamma, \;\; (\eta, \eta_\Gamma) = D^2\mathcal{S}(\bar u_\Gamma)[h_\Gamma, k_\Gamma]\,.$$

Then a straightforward calculation resembling that carried out on page 241 in [6], using the chain rule as the main tool, yields the equality

$$\begin{aligned} D^2\tilde{\mathcal{J}}(\bar u_\Gamma)[h_\Gamma, k_\Gamma] \,&=\, D_{(y,y_\Gamma)}\mathcal{J}((\bar y, \bar y_\Gamma), \bar u_\Gamma)(\eta, \eta_\Gamma) \\ &\quad + D^2\mathcal{J}((\bar y, \bar y_\Gamma), \bar u_\Gamma)[((\varphi, \varphi_\Gamma), h_\Gamma)\,,\, ((\psi, \psi_\Gamma), k_\Gamma)]\,. \end{aligned} \tag{132}$$

For the first summand on the right-hand side of (132), we have

$$\begin{aligned} D_{(y,y_\Gamma)}\mathcal{J}((\bar y, \bar y_\Gamma), \bar u_\Gamma)(\eta, \eta_\Gamma) \,&=\, b_Q \int_0^T \int_\Omega (\bar y - z_Q)\,\eta \,dx\,dt \\ &\quad + b_\Sigma \int_0^T \int_\Gamma (\bar y_\Gamma - z_\Sigma)\,\eta_\Gamma \,d\Gamma dt, \end{aligned} \tag{133}$$

where (η, η_Γ) solves the system (72)–(77). We now claim that

$$\begin{aligned} &b_Q \int_0^T \int_\Omega (\bar y - z_Q)\,\eta \,dx\,dt \,+\, b_\Sigma \int_0^T \int_\Gamma (\bar y_\Gamma - z_\Sigma)\,\eta_\Gamma \,d\Gamma\,dt \\ &= -\int_0^T \int_\Omega f^{(3)}(\bar y)\,\varphi\,\psi\,q \,dx\,dt \,-\, \int_0^T \int_\Gamma f_\Gamma^{(3)}(\bar y_\Gamma)\,\varphi_\Gamma\,\psi_\Gamma\,q_\Gamma \,d\Gamma\,dt\,. \end{aligned} \tag{134}$$

To prove this claim, we test (72) by p, insert $v = \vartheta$ in (123), and add the resulting equations to obtain

$$0 \,=\, \int_0^T \int_\Omega (\partial_t\eta\,p \,+\, q\,\vartheta)\,dx\,dt\,. \tag{135}$$

Next, we test (73) by q. Since $q_{|\Gamma} = q_\Gamma$, we find the identity

$$\begin{aligned}
&\int_0^T\!\!\int_\Omega q\,\vartheta\,dx\,dt = \int_0^T\!\!\int_\Omega \partial_t\eta\,q\,dx\,dt \,+\, \int_0^T\!\!\int_\Omega \nabla\eta\cdot\nabla q\,dx\,dt \\
&+\int_0^T\!\!\int_\Gamma \partial_t\eta_\Gamma\,q_\Gamma\,d\Gamma\,dt \,+\, \int_0^T\!\!\int_\Gamma \nabla_\Gamma\eta_\Gamma\cdot\nabla_\Gamma q_\Gamma\,d\Gamma\,dt \,+\, \int_0^T\!\!\int_\Omega f''(\bar y)\,\eta\,q\,dx\,dt \\
&+\int_0^T\!\!\int_\Omega f^{(3)}(\bar y)\,\varphi\,\psi\,q\,dx\,dt \,+\, \int_0^T\!\!\int_\Gamma f_\Gamma''(\bar y_\Gamma)\,\eta_\Gamma\,q_\Gamma\,d\Gamma\,dt \\
&+\int_0^T\!\!\int_\Gamma f_\Gamma^{(3)}(\bar y_\Gamma)\,\varphi_\Gamma\,\psi_\Gamma\,q_\Gamma\,d\Gamma\,dt\,. \qquad (136)
\end{aligned}$$

Now observe that the initial condition $\eta(0) = \eta_\Gamma(0) = 0$ and the final condition (125) imply, using integration by parts with respect to time, that

$$\begin{aligned}
&\int_0^T\!\!\int_\Omega \partial_t\eta\,(p+q)\,dx\,dt \,+\, \int_0^T\!\!\int_\Gamma \partial_t\eta_\Gamma\,q_\Gamma\,d\Gamma\,dt \\
&= -\int_0^T\!\!\int_\Omega \partial_t(p+q)\,\eta\,dx\,dt \,-\, \int_0^T\!\!\int_\Gamma \eta_\Gamma\,\partial_t q_\Gamma\,d\Gamma\,dt\,.
\end{aligned}$$

Hence, by adding (135) and (136) to each other, we obtain the identity

$$\begin{aligned}
0 = &-\int_0^T\!\!\int_\Omega \partial_t(p+q)\,\eta\,dx\,dt \,-\, \int_0^T\!\!\int_\Gamma \eta_\Gamma\,\partial_t q_\Gamma\,d\Gamma\,dt \,+\, \int_0^T\!\!\int_\Omega \nabla\eta\cdot\nabla q\,dx\,dt \\
&+\int_0^T\!\!\int_\Gamma \nabla_\Gamma\eta_\Gamma\cdot\nabla_\Gamma q_\Gamma\,d\Gamma dt + \int_0^T\!\!\int_\Omega f''(\bar y)\eta\,q\,dx\,dt + \int_0^T\!\!\int_\Omega f^{(3)}(\bar y)\varphi\psi q\,dx\,dt \\
&+\int_0^T\!\!\int_\Gamma f_\Gamma''(\bar y_\Gamma)\,\eta_\Gamma\,q_\Gamma\,d\Gamma\,dt \,+\, \int_0^T\!\!\int_\Gamma f_\Gamma^{(3)}(\bar y_\Gamma)\,\varphi_\Gamma\,\psi_\Gamma\,q_\Gamma\,d\Gamma\,dt\,. \qquad (137)
\end{aligned}$$

Inserting $(v, v_\Gamma) = (\eta, \eta_\Gamma)$ in (124), we finally obtain that

$$\begin{aligned}
0 \,=\, &\int_0^T\!\!\int_\Omega \Big(b_Q(\bar y - z_Q)\,\eta \,+\, f^{(3)}(\bar y)\,\varphi\,\psi\,q\Big)\,dx\,dt \\
&+\int_0^T\!\!\int_\Gamma \Big(b_\Sigma(\bar y_\Gamma - z_\Sigma)\,\eta_\Gamma \,+\, f_\Gamma^{(3)}(\bar y_\Gamma)\,\varphi_\Gamma\,\psi_\Gamma\,q_\Gamma\Big)\,d\Gamma\,dt,
\end{aligned}$$

by comparison. From this the claim (134) follows.

Now we can recall (131)–(134) in order to find the representation formula

$$\begin{aligned}
D^2\tilde J(\bar u_\Gamma)[h_\Gamma, h_\Gamma] \,=\, &b_0\,\|h_\Gamma\|^2_{L^2(\Sigma)} \,+\, \int_0^T\!\!\int_\Omega \Big(b_Q - q\,f^{(3)}(\bar y)\Big)\,|\varphi|^2\,dx\,dt \\
&+\int_0^T\!\!\int_\Gamma \Big(b_\Sigma - q_\Gamma\,f_\Gamma^{(3)}(\bar y_\Gamma)\Big)\,|\varphi_\Gamma|^2\,d\Gamma\,dt\,. \qquad (138)
\end{aligned}$$

Equality (138) gives rise to hope that, under appropriate conditions, $D^2\tilde{\jmath}(\bar{u}_\Gamma)$ might be a positive definite operator on a suitable subset of the space $L^2(\Sigma)$. To formulate such a condition, we introduce for fixed $\tau > 0$ the *set of strongly active constraints for* $\bar{u}_\Gamma$ by

$$A_\tau(\bar{u}_\Gamma) \;:=\; \{(x,t) \in \Sigma : \, |q_\Gamma(x,t) + b_0\, \bar{u}_\Gamma(x,t)| > \tau\}\,, \tag{139}$$

and we define the $\tau-$*critical cone* $C_\tau(\bar{u}_\Gamma)$ to be the set of all $h_\Gamma \in \mathcal{X}_{M_0} := \{h_\Gamma \in \mathcal{X} : \, \|\partial_t h_\Gamma\|_{L^2(\Sigma)} \le M_0\}$ such that

$$h_\Gamma(x,t) \begin{cases} = 0 & \text{if} \quad (x,t) \in A_\tau(\bar{u}_\Gamma) \\ \ge 0 & \text{if} \quad \bar{u}_\Gamma(x,t) = u_{\Gamma,\min} \text{ and } (x,t) \notin A_\tau(\bar{u}_\Gamma) \\ \le 0 & \text{if} \quad \bar{u}_\Gamma(x,t) = u_{\Gamma,\max} \text{ and } (x,t) \notin A_\tau(\bar{u}_\Gamma) \end{cases} . \tag{140}$$

After these preparations, we can formulate the second-order sufficient optimality condition (SSC) as follows.

There exist constants $\delta > 0$ and $\tau > 0$ such that

$$D^2\tilde{\jmath}(\bar{u}_\Gamma)\,[h_\Gamma, h_\Gamma] \;\ge\; \delta\, \|h_\Gamma\|^2_{L^2(\Sigma)} \quad \forall\, h_\Gamma \in C_\tau(\bar{u}_\Gamma),$$

where $D^2\tilde{\jmath}(\bar{u}_\Gamma)\,[h_\Gamma, h_\Gamma]$ is given by (138) with $(\bar{y}, \bar{y}_\Gamma) = \mathcal{S}(\bar{u}_\Gamma)$,

$(\varphi, \varphi_\Gamma) = D\mathcal{S}(\bar{u}_\Gamma)h_\Gamma$ and the associated adjoint state (p, q, q_Γ). (141)

The following result resembles Theorem 5.17 in [6].

Theorem 5. *Suppose that the conditions* **(A1)**–**(A4)** *and* (118) *are fulfilled, and assume* $\bar{u}_\Gamma \in \mathcal{U}_{ad}$, $(\bar{y}, \bar{y}_\Gamma) = \mathcal{S}(\bar{u}_\Gamma)$, *and that the triple* (p, q, q_Γ) *satisfies* (119)–(125). *Moreover, assume that the conditions* (130) *and* (141) *are fulfilled. Then there are constants* $\varepsilon > 0$ *and* $\sigma > 0$ *such that*

$$\tilde{\jmath}(u_\Gamma) \;\ge\; \tilde{\jmath}(\bar{u}_\Gamma) + \sigma\, \|u_\Gamma - \bar{u}_\Gamma\|^2_{L^2(\Sigma)} \quad \textit{for all } u_\Gamma \in \mathcal{U}_{ad} \textit{ with } \|u_\Gamma - \bar{u}_\Gamma\|_{\mathcal{X}} \le \varepsilon\,. \tag{142}$$

In particular, $\bar{u}_\Gamma$ *is locally optimal for* **(CP)** *in the sense of* $\mathcal{X}$.

Proof. The proof closely follows that of [6, Thm. 5.17], and therefore we can refer to [6]. We only indicate one argument that needs additional explanation. To this end, let $u_\Gamma \in \mathcal{U}_{ad}$ be arbitrary. Since $\tilde{\jmath}$ is twice continuously Fréchet differentiable in $\mathcal{U}$, it follows from Taylor's theorem with integral remainder (see, e. g., [4, Thm. 8.14.3, p. 186]) that

$$\tilde{\jmath}(u_\Gamma) - \tilde{\jmath}(\bar{u}_\Gamma) = D\tilde{\jmath}(\bar{u}_\Gamma)v_\Gamma \,+\, \frac{1}{2}\, D^2\tilde{\jmath}(\bar{u}_\Gamma)[v_\Gamma, v_\Gamma] + \, R^{\tilde{\jmath}}(u_\Gamma, \bar{u}_\Gamma)\,, \tag{143}$$

with $v_\Gamma = u_\Gamma - \bar{u}_\Gamma$ and the remainder

$$R^{\tilde{\jmath}}(u_\Gamma, \bar{u}_\Gamma) = \int_0^1 (1-s) \left(D^2 \tilde{\jmath}(\bar{u}_\Gamma + s\, v_\Gamma) - D^2 \tilde{\jmath}(\bar{u}_\Gamma) \right) [v_\Gamma, v_\Gamma] \, ds. \quad (144)$$

Now, we estimate the integrand $(D^2\tilde{\jmath}(\bar{u}_\Gamma + s\, v_\Gamma) - D^2\tilde{\jmath}(\bar{u}_\Gamma))[v_\Gamma, v_\Gamma]$ in (144). To this end, we put

$$(y^s, y^s_\Gamma) = \mathcal{S}(\bar{u}_\Gamma + s v_\Gamma), \;\; (\varphi, \varphi_\Gamma) = D\mathcal{S}(\bar{u}_\Gamma) v_\Gamma, \;\; (\varphi^s, \varphi^s_\Gamma) = D\mathcal{S}(\bar{u}_\Gamma + s v_\Gamma) v_\Gamma,$$
$$(\eta, \eta_\Gamma) = D^2\mathcal{S}(\bar{u}_\Gamma)[v_\Gamma, v_\Gamma], \quad (\eta^s, \eta^s_\Gamma) = D^2\mathcal{S}(\bar{u}_\Gamma + s v_\Gamma)[v_\Gamma, v_\Gamma],$$

and use the representation formulas (131)–(133). We obtain

$$D_{(y,y_\Gamma)}\mathcal{J}((y^s, y^s_\Gamma), \bar{u}_\Gamma + s v_\Gamma)(\eta^s, \eta^s_\Gamma) - D_{(y,y_\Gamma)}\mathcal{J}((\bar{y}, \bar{y}_\Gamma), \bar{u}_\Gamma)(\eta, \eta_\Gamma) = I_1 + I_2, \quad (145)$$

with the integrals

$$I_1 := b_Q \int_0^T \!\! \int_\Omega (y^s - \bar{y})\, \eta \, dx\, dt \, + \, b_\Sigma \int_0^T \!\! \int_\Gamma (y^s_\Gamma - \bar{y}_\Gamma)\, \eta_\Gamma \, d\Gamma \, dt,$$
$$I_2 := b_Q \int_0^T \!\! \int_\Omega (y^s - z_Q)\, (\eta^s - \eta) \, dx\, dt \, + \, b_\Sigma \int_0^T \!\! \int_\Gamma (y^s_\Gamma - z_\Sigma)\, (\eta^s_\Gamma - \eta_\Gamma) \, d\Gamma \, dt. \quad (146)$$

Moreover,

$$\begin{aligned} &D^2\mathcal{J}((y^s, y^s_\Gamma), \bar{u}_\Gamma + s v_\Gamma)[((\varphi^s, \varphi^s_\Gamma), v_\Gamma)\,, \, ((\varphi^s, \varphi^s_\Gamma), v_\Gamma)] \\ &-D^2\mathcal{J}((\bar{y}, \bar{y}_\Gamma), \bar{u}_\Gamma)[((\varphi, \varphi_\Gamma), v_\Gamma)\,, \, ((\varphi, \varphi_\Gamma), v_\Gamma)] \, = \, I_3, \qquad \text{where} \end{aligned}$$

$$I_3 := b_Q \int_0^T \!\! \int_\Omega (\varphi^s - \varphi)(\varphi^s + \varphi) \, dx\, dt \, + \, b_\Sigma \int_0^T \!\! \int_\Gamma (\varphi^s_\Gamma - \varphi_\Gamma)(\varphi^s_\Gamma + \varphi_\Gamma) \, d\Gamma \, dt\,. \quad (147)$$

We now estimate the integrals I_1, I_2, and I_3, where we denote by C_i, $i \in \mathbb{N}$, constants that depend neither on $s \in [0,1]$ nor on $u_\Gamma \in \mathcal{U}_{ad}$. At first, using the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} |I_1| \, &\le \, \max\{b_Q, b_\Sigma\} \, \|(y^s, y^s_\Gamma) - (\bar{y}, \bar{y}_\Gamma)\|_{L^2(0,T;\mathcal{H})} \, \|(\eta, \eta_\Gamma)\|_{L^2(0,T;\mathcal{H})} \\ &\le \, \max\{b_Q, b_\Sigma\} \, \|(y^s, y^s_\Gamma) - (\bar{y}, \bar{y}_\Gamma)\|_{\mathcal{Y}} \, \|(\eta, \eta_\Gamma)\|_{\mathcal{Y}} \\ &\le C_1 \, s \, \|v_\Gamma\|^3_{L^2(\Sigma)}, \end{aligned} \quad (148)$$

where in the last inequality we have employed the estimates (36) and (82). Similarly, we have

$$\begin{aligned} |I_2| &\le \max\{b_Q, b_\Sigma\} \|(y^s, y_\Gamma^s) - (z_Q, z_\Sigma)\|_{L^2(0,T;\mathcal{H})} \|(\eta^s, \eta_\Gamma^s) - (\eta, \eta_\Gamma)\|_{L^2(0,T;\mathcal{H})} \\ &\le \max\{b_Q, b_\Sigma\} \, \|(y^s, y_\Gamma^s) - (z_Q, z_\Sigma)\|_{L^2(0,T;\mathcal{H})} \, \|(\eta^s, \eta_\Gamma^s) - (\eta, \eta_\Gamma)\|_{\mathcal{Y}} \\ &\le C_2 \, s \, \|v_\Gamma\|^3_{L^2(\Sigma)}, \end{aligned} \tag{149}$$

where, for the last inequality, we used **(A1)** and (34) to estimate the first norm and (79) for the second one. Finally, we get

$$\begin{aligned} |I_3| &\le \max\{b_Q, b_\Sigma\} \|(\varphi^s, \varphi_\Gamma^s) - (\varphi, \varphi_\Gamma)\|_{L^2(0,T;\mathcal{H})} \|(\varphi^s, \varphi_\Gamma^s) + (\varphi, \varphi_\Gamma)\|_{L^2(0,T;\mathcal{H})} \\ &\le \max\{b_Q, b_\Sigma\} \, \|(\varphi^s, \varphi_\Gamma^s) - (\varphi, \varphi_\Gamma)\|_{\mathcal{Y}} \, \|(\varphi^s, \varphi_\Gamma^s) + (\varphi, \varphi_\Gamma)\|_{\mathcal{Y}} \\ &\le C_3 \, s \, \|v_\Gamma\|^3_{L^2(\Sigma)}. \end{aligned} \tag{150}$$

For the last inequality, we applied (61) to estimate the first norm and the triangle inequality and (48) to estimate the second one. Combining the above estimates, we thus have finally shown that

$$\left| R^{\tilde{\jmath}}(u_\Gamma, \bar{u}_\Gamma) \right| \;\le\; C_4 \int_0^1 (1-s) \, s \, \|v_\Gamma\|^3_{L^2(\Sigma)} \, ds \;\le\; C_5 \, \|v_\Gamma\|_{\mathcal{X}} \, \|v_\Gamma\|^2_{L^2(\Sigma)}, \tag{151}$$

with global constants $C_4 > 0$ and $C_5 > 0$ that do not depend on the choice of $u_\Gamma \in \mathcal{U}_{ad}$. But this means that

$$\frac{\left| R^{\tilde{\jmath}}(u_\Gamma, \bar{u}_\Gamma) \right|}{\|u_\Gamma - \bar{u}_\Gamma\|^2_{L^2(\Sigma)}} \to 0 \quad \text{as } \|u_\Gamma - \bar{u}_\Gamma\|_{\mathcal{X}} \to 0. \tag{152}$$

With this information at hand, we can argue along exactly the same lines as on pages 292–294 in the proof of Theorem 5.17 in [6] to conclude the validity of the assertion. □

Acknowledgement. P. Colli and G. Gilardi would like to acknowledge some financial support from the MIUR-PRIN Grant 2010A2TFX2 "Calculus of Variations", the GNAMPA (Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni) of INdAM (Istituto Nazionale di Alta Matematica) and the IMATI – C.N.R. Pavia.

References

[1] P. Colli, G. Gilardi and J. Sprekels, On the Cahn–Hilliard equation with dynamic boundary conditions and a dominating boundary potential, *J. Math. Anal. Appl.* 419:972-994, 2014.

[2] P. Colli, G. Gilardi and J. Sprekels, A boundary control problem for the viscous Cahn–Hilliard equation with dynamic boundary conditions, preprint arXiv:1407.3916 [math.AP] 1-27, 2014. Published online 28. April 2015 in *Appl. Math. Optim.*, DOI 10.1007/s00245-015-0299z.

[3] P. Colli and J. Sprekels, Optimal control of an Allen–Cahn equation with singular potentials and dynamic boundary condition, *SIAM J. Control Optim.* 53:213-234, 2015.

[4] J. Dieudonné, Foundations of Modern Analysis, Academic Press, New York, 1960.

[5] M. Heinkenschloss and F. Tröltzsch, Analysis of an SQP method for the control of a phase field equation, *Control Cybernetics* 28:177-211, 1999.

[6] F. Tröltzsch, Optimal Control of Partial Differential Equations: Theory, Methods and Applications, *Graduate Studies in Mathematics* Vol. 112, American Mathematical Society, Providence, Rhode Island, 2010.

Ann. Acad. Rom. Sci.
Ser. Math. Appl.
ISSN 2066-6594 Vol. 7, No. 1/2015

NONLINEAR DELAY EVOLUTION INCLUSIONS WITH GENERAL NONLOCAL INITIAL CONDITIONS *

Mihai Necula† Ioan I. Vrabie‡

Dedicated to the memory of Prof. Dr. Viorel Arnăutu

Abstract

We consider a nonlinear delay differential evolution inclusion subjected to nonlocal implicit initial conditions and we prove an existence result for bounded C^0-solutions.

MSC: 34K09; 34K13; 34K30; 34K40; 35K55; 35L60; 35K91; 47J35

keywords: differential delay evolution inclusion; nonlocal delay initial condition; bounded C^0-solutions; periodic C^0-solutions; anti-periodic C^0-solutions; nonlinear diffusion equation.

1 Introduction

The goal of this paper is to prove an existence result for bounded C^0-solutions to a class of nonlinear delay differential evolution inclusions sub-

*Accepted for publication on December 21-st, 2014

†necula@uaic.ro Department of Mathematics, "Al. I. Cuza" University Iaşi 700506, Romania

‡ivrabie@uaic.ro Department of Mathematics, "Al. I. Cuza" University Iaşi 700506, Romania and Octav Mayer Mathemtics Institute (Romanian Academy) Iaşi 700505, Romania

jected to nonlocal implicit initial conditions of the form

$$\begin{cases} u'(t) \in Au(t) + f(t), & t \in \mathbf{R}_+, \\ f(t) \in F(t, u_t), & t \in \mathbf{R}_+, \\ u(t) = g(u)(t), & t \in [\,-\tau, 0\,], \end{cases} \tag{1}$$

where X is a Banach space, $\tau \geq 0$, $A : D(A) \subseteq X \hookrightarrow X$ is the infinitesimal generator of a nonlinear semigroup of contractions, the multifunction $F : \mathbf{R}_+ \times C([\,-\tau, 0\,]; \overline{D(A)}) \hookrightarrow X$ is nonempty, convex weakly compact valued and strongly-weakly u.s.c., and $g : C_b([\,-\tau, +\infty); \overline{D(A)}) \to C([\,-\tau, 0\,]; \overline{D(A)})$ is nonexpansive and has *affine growth*, i.e. there exists $m_0 \geq 0$ such that

$$\|g(u)\|_{C([\,-\tau,0\,];X)} \leq \|u\|_{C_b([\,0,+\infty);X)} + m_0 \tag{2}$$

for each $u \in C_b([\,-\tau, +\infty); \overline{D(A)})$.

If I is an interval, $C_b(I; X)$ denotes the space of all bounded and continuous functions from I, equipped with the sup-norm $\|\cdot\|_{C_b(I;X)}$, while $C_b(I; \overline{D(A)})$ denotes the closed subset in $C_b(I; X)$ consisting of all elements $u \in C_b(I; X)$ satisfying $u(t) \in \overline{D(A)}$ for each $t \in I$. Let $a \in \mathbf{R}$. On the linear space $C_b([\,a, +\infty); X)$ let us consider the family of seminorms $\{\|\cdot\|_k;\ k \in \mathbf{N},\ k \geq a\}$, defined by $\|u\|_k = \sup\{\|u(t)\|;\ t \in [\,a, k\,]\}$ for each $k \in \mathbf{N}$, $k \geq a$. Endowed with this family of seminorms, $C_b([\,a, +\infty); X)$ is a separated locally convex space, denoted by $\widetilde{C}_b([\,a, +\infty); X)$. Further, $C([\,a, b\,]; X)$ stands for the space of all continuous functions from $[\,a, b\,]$ to X endowed with the sup-norm $\|\cdot\|_{C([\,a,b\,];X)}$ and $C([\,a, b\,]; \overline{D(A)})$ is the closed subset of $C([\,a, b\,]; X)$ containing all $u \in C([\,a, b\,]; X)$ with $u(t) \in \overline{D(A)}$ for each $t \in [\,a, b\,]$. Finally, if $u \in C_b([\,-\tau, +\infty); X)$ and $t \in \mathbf{R}_+$, $u_t \in C([\,-\tau, 0\,]; X)$ is defined by

$$u_t(s) := u(t + s)$$

for each $s \in [\,-\tau, 0\,]$.

The existence problem on the standard compact interval $[\,0, 2\pi\,]$, in the simplest case when $\tau = 0$, i.e. when the delay is absent, was studied by Paicu, Vrabie [41]. In this case $C([\,-\tau, 0\,]; \overline{D(A)})$ identifies with $\overline{D(A)}$, F identifies with a multifunction from $[\,0, 2\pi\,] \times X$ to X. By using an interplay between compactness arguments and invariance techniques, they have proved an existence result handling periodic, anti-periodic, mean-value evolution inclusions subjected to initial condition expressed by an integral with

respect to a Radon measure μ. A very important specific case concerns T-periodic problems, which corresponds to the choice of g as $g(u) = u(T)$, was studied by Paicu [39]. For F single-valued, this case was analyzed by Aizicovici, Papageorgiou, Staicu [3], Caşcaval, Vrabie [18], Hirano, Shioji [34], Paicu [40], Vrabie [44]. For a survey concerning: periodic, anti-periodic, quasi-periodic and almost periodic solutions to differential inclusions, see Andres [6]. As long as differential inclusions subjected to general nonlocal initial conditions without delay are concerned, we mention the papers of Aizicovici, Staicu [5] and Paicu, Vrabie [41]. The case of periodic retarded equations and inclusions subjected to nonlocal initial conditions were studied by Vrabie [46], and Chen, Wang, Zhou [20], while the general delay equations was considered by Burlică, Roşu [14] and Vrabie [48], [49] and [50].

Existence results in the periodic abstract undelayed case were obtained by Aizicovici, Papageorgiou, Staicu [3], Caşcaval, Vrabie [18], Hirano, Shioji [34], Paicu [40], Vrabie [44], while the anti-periodic case was considered by Aizicovici, Pavel, Vrabie [4]. The semilinear case of undelayed differential equations subjected to nonlocal initial data, was initiated by the pioneering work of Byszewski [15]. Further steps in this direction were made by Byszewski [16], Byszewski, Lakshmikantham [17], Aizicovici, Lee [1], Aizicovici, McKibben [2], Zhenbin Fan, Qixiang Dong, Gang Li [27], García-Falset [29] and García-Falset, Reich [30]. All these studies are strongly motivated by the fact that specific problems of this kind describe the evolution of various phenomena in Physics, Meteorology, Thermodynamics, Population Dynamics. A model of the gas flow through a thin transparent tube, expressed as a problem with nonlocal initial conditions, was analyzed in Deng [24]. Some models in Pharmacokinetics were discussed in the monograph of McKibben [35, Section 10.2, pp. 394–398]. Models arising from Physics were analyzed by Olmstead, Roberts [38] and Shelukhin [43]. Linear second order evolution equations subjected to linear nonlocal initial conditions in Hilbert triples were considered in Avalishvili, Avalishvili [8] and motivated by mathematical models for long-term reliable weather forecasting as mentioned in Rabier, Courtier, Ehrendorfer [42]. For Navier-Stokes equations subjected to initial nonlocal conditions see Gordeziani [32]. Classical nonlinear delay evolution initial-value problems, i.e. when $g \equiv \psi$ with $\psi \in C([\tau, 0]; \overline{D(A)})$, were considered by Mitidieri, Vrabie [36] and [37], also by using compactness arguments. It should be emphasized that in Mitidieri, Vrabie [36] and [37] the general assumptions on the forcing term F are very general allowing – in certain specific cases when A is a second order elliptic operator – F to depend on Au as well.

Our paper extends the main result in Vrabie [47] to cover the more general case in which g has affine rather than linear growth. This case is important in applications and does not follow by a simple modification of the arguments used in Vrabie [47].

The paper is divided into 7 sections. In Section 2 we have included some concepts and results widely used subsequently. In Section 3 we prove an existence and uniqueness result for the unperturbed problem (1) which, although auxiliary, is important by its own. Section 4 collects the hypotheses used and provides some comments on several remarkable particular cases handled by the general frame considered. Section 5 is devoted to the statement of the main result, i.e. Theorem 7 and to a short description of the idea of the proof. Section 6 is concerned with the proof of the main result and the last Section 7 contains an example illustrating the possibilities of the abstract developed theory.

2 Preliminaries

Although the paper is almost self-contained, some familiarity with the basic concepts and results on nonlinear evolution equations governed by m-dissipative operators, delay evolution equations and on multifunction theory would be welcome. For details in these three topics, we refer the reader, in order, to Barbu [11], Hale [33] and Vrabie [45]. However, we recall for easy reference the most important notions and results we will use in the sequel.

Definition 1 If X is a Banach space and $\mathcal{C} \subseteq X$, the multifunction $F : \mathcal{C} \hookrightarrow X$ is said (*strongly-weakly*) *upper semicontinuous* (*u.s.c.*) *at* $\xi \in \mathcal{C}$ if for every (weakly) open neighborhood V of $F(\xi)$ there exists an open neighborhood U of ξ such that $F(\eta) \subseteq V$ for each $\eta \in U \cap \mathcal{C}$. We say that F is (*strongly-weakly*) *u.s.c. on* $\mathcal{C}$ if it is (strongly-weakly) u.s.c. at each $\xi \in \mathcal{C}$.

Definition 2 A multifunction $F : I \times \mathcal{C} \hookrightarrow X$ is said to be *almost strongly-weakly u.s.c.* if for each $\gamma > 0$ there exists a Lebesgue measurable subset $E_\gamma \subseteq I$ whose Lebesgue measure $\lambda(E_\gamma) \leq \gamma$ and such that F it is strongly-weakly u.s.c. from $(I \setminus E_\gamma) \times \mathcal{C}$ to X.

Remark 1 If the sequence $(\varepsilon_n)_n$ is strictly decreasing to 0, we can always choose the sequence $(E_{\varepsilon_n})_n$, where E_{ε_n} corresponds to ε_n as specified in Definition 2, such that $E_{\varepsilon_{n+1}} \subseteq E_{\varepsilon_n}$, for $n = 0, 1, \ldots$.

We also need the following general fixed point theorem for multifunctions obtained independently by Ky Fan [28] and Glicksberg [31].

Theorem 1 (Ky Fan-Glicksberg) *Let K be a nonempty, convex and compact set in a separated locally convex space and let $\Gamma : K \hookrightarrow K$ be a nonempty, closed and convex valued multifunction with closed graph. Then Γ has at least one fixed point, i.e. there exists $f \in K$ such that $f \in \Gamma(f)$.*

A very useful variant of Theorem 1, is

Theorem 2 *Let K be a nonempty, convex and closed set in a separated locally convex space and let $\Gamma : K \hookrightarrow K$ be a nonempty, closed and convex valued multifunction with closed graph. If $\Gamma(K) := \cup_{x \in K} \Gamma(x)$ is relatively compact, then Γ has at least one fixed point, i.e. there exists $f \in K$ such that $f \in \Gamma(f)$.*

Proof. Since K is closed, convex and $\Gamma(K) \subseteq K$, we have

$$\overline{\operatorname{conv} \Gamma(K)} \subseteq \overline{\operatorname{conv} K} = K.$$

So,

$$\Gamma(\overline{\operatorname{conv} \Gamma(K)}) \subseteq \Gamma(K) \subseteq \overline{\operatorname{conv} \Gamma(K)},$$

which shows that the set $\mathcal{C} := \overline{\operatorname{conv} \Gamma(K)}$, which by Mazur's Theorem, i.e. Dunford, Schwartz [22, Theorem 6, p. 416] is compact, is nonempty, closed, convex and $\Gamma(\mathcal{C}) \subseteq \mathcal{C}$. So, we are in the hypotheses of Theorem 1, with K substituted by $\mathcal{C} \subseteq K$, wherefrom the conclusion. □

Since, by Edwards [23, Theorem 8.12.1, p. 549], the weak closure of a weakly relatively compact set, in a Banach space, coincides with its weak sequential closure, Theorem 2 implies:

Theorem 3 *Let K be a nonempty, convex and weakly compact set in Banach space and let $\Gamma : K \hookrightarrow K$ be a nonempty, closed and convex valued multifunction with sequentially closed graph. Then Γ has at least one fixed point, i.e. there exists $f \in K$ such that $f \in \Gamma(f)$.*

In the single-valued case, Theorem 3 is due to Arino, Gautier, Penot [7].

If $x, y \in X$, we denote by $[\, x, y \,]_{\pm}$ the *right* (*left*) *directional derivative of the norm* calculated at x in the direction y, i.e.

$$[\, x, y \,]_{+} = \lim_{h \downarrow 0} \frac{\|x + hy\| - \|x\|}{h} \qquad \left([\, x, y \,]_{-} = \lim_{h \uparrow 0} \frac{\|x + hy\| - \|x\|}{h} \right).$$

We recall that:

$$[\, x, y + ax \,]_{\pm} = [\, x, y \,]_{\pm} + a\|x\| \tag{3}$$

for $a \in \mathbf{R}$. See Barbu [11, Proposition 3.7, p. 101].

We say that the operator $A : D(A) \subseteq X \hookrightarrow X$ is *dissipative* if

$$[\, x_1 - x_2, y_1 - y_2 \,]_- \leq 0$$

for each $x_i \in D(A)$ and $y_i \in Ax_i$, $i = 1, 2$, and *m-dissipative* if it is dissipative and, for each $\lambda > 0$, or equivalently for some $\lambda > 0$, $R(I - \lambda A) = X$.

Let $A : D(A) \subseteq X \hookrightarrow X$ be an m-dissipative operator, let $\xi \in \overline{D(A)}$, $f \in L^1(a, b; X)$ and let us consider the differential equation

$$u'(t) \in Au(t) + f(t). \tag{4}$$

Theorem 4 (Benilan) *Let $\omega \in \mathbf{R}$ and let $A : D(A) \subseteq X \hookrightarrow X$ be an m-dissipative operator such that $A+\omega I$ is dissipative. Then, for each $\xi \in \overline{D(A)}$ and $f \in L^1(a, b\,; X)$, there exists a unique C^0-solution of (4) on $[\, a, b \,]$ which satisfies $u(a) = \xi$. Furthermore, if $f, g \in L^1(a, b\,; X)$ and u, v are the two C^0-solutions of (4) corresponding to f and g respectively, then:*

$$\|u(t) - v(t)\| \leq e^{-\omega(t-s)}\|u(s) - v(s)\| + \int_s^t e^{-\omega(t-\theta)}\|f(\theta) - g(\theta)\|d\theta \tag{5}$$

for each $a \leq s \leq t \leq b$.

See Benilan [12], or Barbu [11, Theorem 4.1, p. 128].

We denote by $u(\cdot, a, \xi, f)$ the unique C^0-solution of the problem (4) satisfying

$$u(a, a, \xi, f) = \xi$$

and we notice that $u(t, 0, \xi, 0) = S(t)\xi$, where $\{S(t);\ S(t) : \overline{D(A)} \to \overline{D(A)}\}$ is the semigroup of nonexpansive mappings generated by A via the Crandall-Liggett Exponential Formula. See Crandall, Liggett [21].

We recall that the semigroup $\{S(t);\ S(t) : \overline{D(A)} \to \overline{D(A)}\}$ is called *compact* if, for each $t > 0$, $S(t)$ is a compact operator.

We conclude this section with some compactness results concerning the set of C^0-solutions of the problem (4) whose initial data $u(a)$ and forcing terms f belong to some subsets B, in $\overline{D(A)}$, and respectively $\mathcal{F}$, in $L^1(a, b; X)$. First, we introduce:

Definition 3 Let (Ω, Σ, μ) be a complete measure space, $\mu(\Omega) < +\infty$. A subset $\mathcal{F} \subseteq L^1(\Omega, \mu; X)$ is called *uniformly integrable* if for each $\varepsilon > 0$ there exists $\delta(\varepsilon) > 0$ such that

$$\int_E \|f(t)\|\, d\mu(t) \leq \varepsilon$$

for each $f \in \mathcal{F}$ and each $E \in \Sigma$ satisfying $\mu(E) \leq \delta(\varepsilon)$.

The next result is an extension of a compactness theorem due to Baras [10].

Theorem 5 *Let X be a Banach space, let $A : D(A) \subseteq X \hookrightarrow X$ be an m-dissipative operator and let us assume that A generates a compact semigroup. Let $B \subseteq \overline{D(A)}$ be bounded and let $\mathcal{F}$ be uniformly integrable in $L^1(a,b;X)$. Then, for each $\sigma \in (a,b)$, the set $\{u(\cdot,a,\xi,f)\,;\ (\xi,f) \in B \times \mathcal{F}\}$ is relatively compact in $C([\,\sigma,b\,];X)$. If, in addition, B is relatively compact, then $\{u(\cdot,a,\xi,f)\,;\ (\xi,f) \in B \times \mathcal{F}\}$ is relatively compact even in $C([\,a,b\,];X)$.*

See Vrabie [45, Theorems 2.3.2 and 2.3.3, pp. 46–47].

Definition 4 An m-dissipative operator A is called of *complete continuous type* if for each $a < b$ and each sequences $(f_n)_n$ in $L^1(a,b;X)$ and $(u_n)_n$ in $C([\,a,b\,];X)$, with u_m a C^0-solution on $[\,a,b\,]$ of the problem $u'_m(t) \in Au_m(t) + f_m(t), \quad m = 1,2,\ldots$ satisfying:

$$\begin{cases} \lim\limits_n f_n = f & \text{weakly in} \quad L^1(a,b;X), \\ \lim\limits_n u_n = u & \text{strongly in} \quad C([\,a,b\,];X), \end{cases}$$

it follows that u is a C^0 solution on $[\,a,b\,]$ of the limit problem $u'(t) \in Au(t) + f(t)$.

Remark 2 If the topological dual of X is uniformly convex and A generates a compact semigroup, then A is of complete continuous type. See Vrabie [45, Corollary 2.3.1, p. 49]. An m-dissipative operator of complete continuous type in a nonreflexive Banach space (and, by consequence, whose dual is not uniformly convex) is the nonlinear diffusion operator $\Delta\varphi$ in $L^1(\Omega)$. See the example below.

Example 1 Let Δ be the Laplace operator in the sense of distributions over Ω. Let $\varphi : D(\varphi) \subseteq \mathbf{R} \hookrightarrow \mathbf{R}$, let $u : \Omega \to D(\varphi)$ and let us denote by

$$\mathcal{S}_\varphi(u) = \{v \in L^1(\Omega);\ v(x) \in \varphi(u(x)), \quad \text{a.e. for } x \in \Omega\}.$$

We recall that $\varphi : D(\varphi) \subseteq \mathbf{R} \hookrightarrow \mathbf{R}$ is said to be *maximal monotone* if $-\varphi$ is m-dissipative.

The (i) part in Theorem 6 below is due to Brezis, Strauss [13], the (ii) part to Badii, Díaz, Tesei [9] and the (iii) part to Cârjă, Necula, Vrabie [19].

Theorem 6 *Let Ω be a nonempty, bounded and open subset in $\mathbf{R}^d$ with C^1 boundary Σ and let $\varphi : D(\varphi) \subseteq \mathbf{R} \hookrightarrow \mathbf{R}$ be maximal monotone with $0 \in \varphi(0)$.*

(*i*) *Then the operator* $\Delta\varphi : D(\Delta\varphi) \subseteq L^1(\Omega) \hookrightarrow L^1(\Omega)$, *defined by*

$$\begin{cases} D(\Delta\varphi) = \{u \in L^1(\Omega);\ \exists v \in \mathcal{S}_\varphi(u) \cap W_0^{1,1}(\Omega),\ \Delta v \in L^1(\Omega)\} \\ \Delta\varphi(u) = \{\Delta v;\ v \in \mathcal{S}_\varphi(u) \cap W_0^{1,1}(\Omega)\} \cap L^1(\Omega) \quad \text{for } u \in D(\Delta\varphi), \end{cases}$$

is m-dissipative on $L^1(\Omega)$.

(*ii*) *If, in addition,* $\varphi : \mathbf{R} \to \mathbf{R}$ *is continuous on* $\mathbf{R}$ *and* C^1 *on* $\mathbf{R} \setminus \{0\}$ *and there exist two constants* $C > 0$ *and* $\alpha > 0$ *if* $d \leq 2$ *and* $\alpha > (d-2)/d$ *if* $d \geq 3$ *such that*

$$\varphi'(r) \geq C|r|^{\alpha-1}$$

for each $r \in \mathbf{R} \setminus \{0\}$, *then* $\Delta\varphi$ *generates a compact semigroup.*

(*iii*) *In the hypotheses of* (*ii*), $\Delta\varphi$ *is of complete continuous type.*

For the proof of (*i*) see Barbu [11, Theorem 3.5, p. 115], for the proof of (*ii*) see Vrabie [45, Theorem 2.7.1, p. 70] and for proof of the (*iii*) – which rests heavily on slight extension of a continuity result established in Díaz, Vrabie [26, Corollary 3.1, p. 527] which, in turn, follows from a compactness result due to Díaz, Vrabie [25] –, see Cârjă, Necula, Vrabie [19, Theorem 1.7.9, p. 22].

3 An auxiliary lemma

We begin by considering the problem

$$\begin{cases} u'(t) \in Au(t) + f(t), & t \in \mathbf{R}_+, \\ u(t) = g(u)(t), & t \in [\,-\tau, 0\,]. \end{cases} \tag{6}$$

Lemma 1 *Let us assume that* A *is m-dissipative,* $0 \in D(A)$, $0 \in A0$ *and there exists* $\omega > 0$ *such that* $A + \omega I$ *is dissipative, too. Let us assume, in addition, that there exists* $a > 0$ *such that* $g : C_b([\,-\tau, +\infty); \overline{D(A)}) \to C([\,-\tau, 0\,]; \overline{D(A)})$ *satisfies*

$$\|g(v) - g(\widetilde{v})\|_{C_b([\,-\tau,0\,];\overline{D(A)})} \leq \|v - \widetilde{v}\|_{C_b([\,a,+\infty);\overline{D(A)})}, \tag{7}$$

for each $v, \widetilde{v} \in C_b([\,-\tau, +\infty); \overline{D(A)})$ *and has affine growth, i.e. satisfies* (2). *Then, for each* $f \in L^\infty(\mathbf{R}_+; X) \cap L^1(\mathbf{R}_+; X)$, (6) *has a unique* C^0*-solution* $u \in C_b([\,-\tau, +\infty); \overline{D(A)})$.

Remark 3 If $g : C_b([\,-\tau,+\infty);\overline{D(A)}) \to C([\,-\tau,0\,];\overline{D(A)})$ satisfies (7), then g depends only on the restriction $v_{|[\,a,+\infty)}$ of v to $[\,a,+\infty)$.

We can now pass to the proof of Lemma 1.

Proof. Let us observe first that, for each $v \in C_b([\,-\tau,+\infty);\overline{D(A)})$, the initial value problem for the delay equation

$$\begin{cases} u'(t) \in Au(t) + f(t), & t \in \mathbf{R}_+, \\ u(t) = g(v)(t), & t \in [\,-\tau,0\,] \end{cases} \tag{8}$$

has a unique C^0-solution $u : [\,-\tau,+\infty) \to \overline{D(A)}$. Clearly, u is bounded on $[\,-\tau,0\,]$ because it is continuous. Next, recalling that $0 \in A0$, from Theorem 4 we conclude that

$$\|u(t)\| \le e^{-\omega t}\|u(0)\| + \int_0^t e^{-\omega(t-\theta)}\|f(\theta)\|d\theta$$

$$\le \|u(0)\| + \frac{1}{\omega}\|f\|_{L^\infty(\mathbf{R}_+;X)},$$

for each $t \ge 0$. Finally, since u is bounded on both $[\,-\tau,0\,]$ and $[\,0,+\infty)$, it follows that $u \in C_b([\,-\tau,+\infty);\overline{D(A)})$.

Now let us observe that, in view of Remark 3, $g(v)(t) = g(\widetilde{v})(t)$ for each $t \in [\,-\tau,0\,]$ whenever v and $\widetilde{v}$ coincide on $[\,a,+\infty)$ and so, g depends only on the restriction of v on $[\,a,+\infty)$ To conclude the proof, it suffices to show that the operator

$$Q : C_b([\,a,+\infty);\overline{D(A)}) \to C_b([\,a,+\infty);\overline{D(A)}),$$

defined by

$$Q(v) := u_{|[\,a,+\infty)},$$

where u is the unique C^0-solution of the problem (8), is a strict contraction. Hence by the Banach Fixed Point Theorem, Q has a unique fixed point $v = u_{|[\,a,+\infty)}$ and

$$u(t) = \begin{cases} u(t), & t \in \mathbf{R}_+ \\ g(v)(t), & t \in [\,-\tau,0\,], \end{cases}$$

is the unique C^0-solution of (6).

To this end, let $v, \widetilde{v} \in C_b([\,a,+\infty);\overline{D(A)})$ and $t \in [\,a,+\infty)$ be arbitrary. We have

$$\|Q(v)(t) - Q(\widetilde{v})(t)\| \le e^{-\omega t}\|Q(v)(0) - Q(\widetilde{v})(0)\|$$

$$\leq e^{-\omega a}\|g(v)(0)-g(\widetilde{v})(0)\| \leq e^{-\omega a}\|v-\widetilde{v}\|_{C_b([\,a,+\infty);X)}.$$

To complete the proof, we have merely to observe that

$$\|Q(v)-Q(\widetilde{v})\|_{C_b([\,a,+\infty);X)} \leq e^{-\omega a}\|v-\widetilde{v}\|_{C_b([\,a,+\infty);X)}$$

for each $v, \widetilde{v} \in C_b([\,a,+\infty);\overline{D(A)})$. □

4 The general frame and basic assumptions

In the sequel we shall denote by $z : [\,-\tau,+\infty) \to \overline{D(A)}$ the unique C^0-solution of the unperturbed problem

$$\begin{cases} z'(t) \in Az(t), & t \in \mathbf{R}_+, \\ z(t) = g(z)(t), & t \in [\,-\tau,0\,]. \end{cases} \tag{9}$$

which, in view of Lemma 1, belongs to $C_b([\,-\tau,+\infty);\overline{D(A)})$.

The assumptions we need in that follows are listed below.

(H_A) $A : D(A) \subseteq X \hookrightarrow X$ is an operator with the properties:

- (A_1) A is m-dissipative, there exists $\omega > 0$ such that $A+\omega I$ is dissipative too, $0 \in D(A)$, $0 \in A0$ and $\overline{D(A)}$ is convex;
- (A_2) the semigroup generated by A on $\overline{D(A)}$ is compact;
- (A_3) A is of complete continuous type. See Definition 4.

(H_F) $F : \mathbf{R}_+ \times C([\,-\tau,0\,];\overline{D(A)}) \hookrightarrow X$ is a nonempty, convex and weakly compact valued, almost strongly-weakly u.s.c. multifunction. See Definition 2.

(H_I) There exists $r > 0$ such that for each $t \in \mathbf{R}_+$, each $v \in C([\,-\tau,0\,];\overline{D(A)})$, with $\|v-z_t\|_{C([\,-\tau,0\,];X)} = r$ and $f \in F(t,v)$, we have $[\,v(0)-z(t),f\,]_+ \leq 0$, where z is the unique C^0-solution of the unperturbed problem (9).

(H_I') There exists $r > 0$ such that for each $t \in \mathbf{R}_+$, each $v \in C([\,-\tau,0\,];\overline{D(A)})$ with $\|v(0)-z(t)\| > r$ and $f \in F(t,v)$, we have $[\,v(0)-z(t),f\,]_+ \leq 0$, where z is the unique C^0-solution of the unperturbed problem (9).

(H_B) There exists $\ell \in L^\infty(\mathbf{R}_+;\mathbf{R}_+) \cap L^1(\mathbf{R}_+;\mathbf{R}_+)$ such that for almost every $t \in \mathbf{R}_+$ and for each $v \in C([\,-\tau,0\,];\overline{D(A)})$ satisfying $\|v(0)-z(t)\| \leq r$, where $r > 0$ is given by (H_I), and each $f \in F(t,v)$, we have

$$\|f\| \leq \ell(t).$$

(H'_B) There exists $\ell \in L^\infty(\mathbf{R}_+;\mathbf{R}_+) \cap L^1(\mathbf{R}_+;\mathbf{R}_+)$ such that

$$\|f\| \le \ell(t)$$

for each $v \in C([\,-\tau,0\,];\overline{D(A)})$, each $f \in F(t,v)$ and a.e. for $t \in \mathbf{R}_+$.

(H_g) $g : C_b([\,-\tau,+\infty);\overline{D(A)}) \to C([\,-\tau,0\,];\overline{D(A)})$ satisfies:

- (g_1) g has affine growth, i.e. there exists $m_0 \ge 0$ such that for each u in $C_b([\,-\tau,+\infty);\overline{D(A)})$, g satisfies (2);
- (g_2) there exists $a > 0$ such that for each $u, v \in C_b([\,-\tau,+\infty);\overline{D(A)})$, we have

$$\|g(u) - g(v)\|_{C([\,-\tau,0\,];X)} \le \|u - v\|_{C_b([\,a,+\infty);X)}\,;$$

- (g_4) g is continuous from $\widetilde{C}_b([\,-\tau,+\infty);\overline{D(A)})$ to $C([\,-\tau,0\,];\overline{D(A)})$.

Remark 4 The hypothesis (H_I) ensures the invariance of a certain moving set with respect to the C^0-solutions of the problem

$$\begin{cases} u'(t) \in Au(t) + f(t), & t \in \mathbf{R}_+, \\ u(t) = g(v)(t), & t \in [\,-\tau,0\,]. \end{cases}$$

Namely, if a C^0-solution u of the problem above satisfies the initial constraint $u(t) - z(t) \in D(0,r)$ for each $t \in [\,-\tau,0\,]$, where z is the unique C^0-solution of (9), then (H_I) implies that u satisfies the very same constraint for all t belonging to domain of existence of u.

If $\|g(u)\|_{C([\,-\tau,0\,];X)} \le \|u\|_{C_b([\,0,+\infty);X)}$ for each $u \in C_b([\,-\tau,+\infty);X)$, case in which we will say that g *has linear growth*, we have $g(0) = 0$ and, accordingly, the unique C^0-solution z of (9) is identically 0. So, in this case, the invariance condition is nothing but a variant of the condition (H_3) in Vrabie [47].

Conditions $(g_1) \sim (g_2)$ and (g_4) are satisfied by all functions g of the general form specified in Remark 5 below.

Remark 5 Let $0 \le \tau < T$. If the function g is defined as

- *(i)* $g(u)(t) = u(T+t)$, $t \in [\,-\tau,0\,]$ (T-periodicity condition);
- *(ii)* $g(u)(t) = -u(T+t)$, $t \in [\,-\tau,0\,]$ (T-antiperiodicity condition);

(*iii*) $g(u)(t) = \int_{\tau}^{+\infty} k(\theta)u(t+\theta)\,d\theta$, $t \in [-\tau, 0]$, where $k \in L^1([\tau, +\infty); \mathbf{R})$ and $\int_{\tau}^{+\infty} |k(\theta)|\,d\theta = 1$ (mean condition);

(*iv*) $g(u)(t) = \sum_{i=1}^{n} \alpha_i u(t+t_i)$ for each $t \in [-\tau, 0]$, where $\sum_{i=1}^{n} |\alpha_i| \leq 1$ and $\tau < t_1 < t_2 < \cdots < t_n = T$ are arbitrary, but fixed (multi-point discrete mean condition);

then g satisfies (g_1) with $m_0 = 0$ and (g_2) with $a = T - \tau > 0$. A more general case is that in which the support of the measure μ is in $(\tau, +\infty)$ and the function is g given by

$$g(u)(t) = \int_{\tau}^{+\infty} \mathcal{N}(u(t+\theta))\,d\mu(\theta) + \psi(t), \tag{10}$$

for each $u \in C_b([-\tau, +\infty); \overline{D(A)})$ and $t \in [-\tau, 0]$. Here $\mathcal{N} : X \to X$ is a (possible nonlinear) nonexpansive operator with $\mathcal{N}(0) = 0$ and μ is a σ-finite and complete measure on $[\tau, +\infty)$, for which there exists $b > \tau$ such that supp $\mu = [b, +\infty)$, $\mu([b, +\infty)) = 1$ and $\psi \in C([-\tau, 0]; X)$ is such that $g(u)(t) \in \overline{D(A)}$ for each $t \in [-\tau, 0]$. Obviously, in this case, the constant $a > 0$ in (g_2) is exactly $b - \tau$.

Remark 6 From (g_2), (g_4) and Remark 3, we conclude that, for each convergent sequence $(u_k)_k$ in $\widetilde{C}_b([a, +\infty); \overline{D(A)})$ to some limit u we have $\lim_k g(u_k) = g(u)$ in $C([-\tau, 0]; X)$.

5 The main result

We may now proceed to the statement of the main result in this paper.

Theorem 7 *If (H_A), (H_F), (H_I), (H_B) and (H_g) are satisfied, then (1) has at least one C^0-solution, $u \in C_b([-\tau, +\infty); \overline{D(A)})$ satisfying $u(t) - z(t) \in D(0, r)$ for each $t \in \mathbf{R}_+$, where z is the unique C^0-solution of (9) and $r > 0$ is given by (H_I).*

We will prove our Theorem 7 with the help of:

Theorem 8 *If (H_A), (H_F), (H_I'), (H_B') and (H_g) are satisfied, then (1) has at least one C^0-solution, $u \in C_b([-\tau, +\infty); \overline{D(A)})$ and $u(t) - z(t) \in D(0, r)$ for each $t \in \mathbf{R}_+$, where z is the unique C^0-solution of (9) and $r > 0$ is given by (H_I').*

The proof of Theorem 8 is divided into four steps.

The first step. We begin by showing that, for each $\varepsilon \in (0,1)$ and $f \in L^1(\mathbf{R}_+; X)$, the problem

$$\begin{cases} u'(t) \in Au(t) - \varepsilon[\, u(t) - z(t)\,] + f(t), & t \in \mathbf{R}_+, \\ u(t) = g(u)(t), & t \in [\,-\tau, 0\,], \end{cases} \tag{11}$$

has a unique C^0-solution $u_\varepsilon^f \in C_b([\,-\tau, +\infty); \overline{D(A)})$.

The second step. We show that for each fixed $\varepsilon \in (0,1)$, the operator $f \mapsto u_\varepsilon^f$, which associates to f the unique C^0-solution u_ε^f of the problem (11), is compact from $L^\infty(\mathbf{R}_+; X) \cap L^1(\mathbf{R}_+; X)$ to $\widetilde{C}_b([\,-\tau, +\infty); \overline{D(A)})$.

The third step. As F is almost strongly-weakly u.s.c. – see Definition 1 –, it follows that, for the very same $\varepsilon > 0$, there exists $E_\varepsilon \subseteq \mathbf{R}_+$ whose Lebesgue measure $\lambda(E_\varepsilon) \leq \varepsilon$ and such that $F_{|(\mathbf{R}_+ \setminus E_\varepsilon) \times C([\,-\tau,0\,]; \overline{D(A)})}$ is strongly-weakly u.s.c., we construct an approximation for F as follows. Let

$$\begin{aligned} D(F) &= \mathbf{R}_+ \times C([\,-\tau, 0\,]; \overline{D(A)}), \\ D_\varepsilon(F) &= (\mathbf{R}_+ \setminus E_\varepsilon) \times C([\,-\tau, 0\,]; \overline{D(A)}) \end{aligned}$$

and let us define the multifunction $F_\varepsilon : \mathbf{R}_+ \times C([\,-\tau, 0\,]; \overline{D(A)}) \hookrightarrow X$, by

$$F_\varepsilon(t, v) = \begin{cases} F(t, v), & (t, v) \in D_\varepsilon(F), \\ \{0\}, & (t, v) \in D(F) \setminus D_\varepsilon(F). \end{cases} \tag{12}$$

Further, we prove that the multifunction $f \mapsto \operatorname{Sel} F_\varepsilon(\cdot, u_{\varepsilon\,(\cdot)}^f)$, where

$$\operatorname{Sel} F_\varepsilon(\cdot, u_{\varepsilon\,(\cdot)}^f) = \{h \in L^1(\mathbf{R}_+; X);\ h(t) \in F_\varepsilon(t, u_{\varepsilon\,t}^f) \text{ a.e. } t \in \mathbf{R}_+\},$$

maps some nonempty, convex and weakly compact set $\mathcal{K} \subseteq L^1(\mathbf{R}_+; X)$ into itself and has weakly$\times$weakly sequentially closed graph. Then, we are in the hypotheses of Theorem 3, wherefrom it follows that this mapping has at least one fixed point which, by means of $f \mapsto u_\varepsilon^f$, produces a C^0-solution for the approximate problem

$$\begin{cases} u'(t) \in Au(t) - \varepsilon[\, u(t) - z(t)\,] + f(t), & t \in \mathbf{R}_+, \\ f(t) \in F_\varepsilon(t, u_t), & t \in \mathbf{R}_+, \\ u(t) = g(u)(t), & t \in [\,-\tau, 0\,], \end{cases} \tag{13}$$

where F_ε is defined by (12).

The fourth step. For each $\varepsilon \in (0,1)$, we fix a C^0-solution u_ε of the problem (13), and we show that there exists a sequence $\varepsilon_n \downarrow 0$ such that $(u_{\varepsilon_n})_n$ converges in $\widetilde{C}_b([\,0, +\infty); \overline{D(A)})$ to a C^0-solution of the problem (1).

6 Proofs of Theorems 7 and 8

We begin with the proofs of the four steps outlined above which are labeled here as four lemmas.

Lemma 2 *Let us assume that* (A_1) *in* (H_A)*, and* $(g_1) \sim (g_2)$ *in* (H_g) *are satisfied. Then, for each* $\varepsilon > 0$ *and each* $f \in L^\infty(\mathbf{R}_+; X) \cap L^1(\mathbf{R}_+; X)$*, the problem* (11) *has a unique* C^0*-solution* $u_\varepsilon^f : [\,-\tau, +\infty) \to X$ *which belongs to* $C_b([\,-\tau, +\infty); \overline{D(A)})$*. Moreover,* u_ε^f *satisfies*

$$\|u_\varepsilon^f - z\|_{C_b([\,-\tau,+\infty);X)} \leq \frac{1}{\varepsilon}\|f\|_{L^\infty(\mathbf{R}_+;X)}, \tag{14}$$

where z *is the unique* C^0*-solution of the problem* (9)*.*

Proof. First, let us observe that the problem (11) has the form

$$\begin{cases} u'(t) \in A_\varepsilon u(t) + f_\varepsilon(t), & t \in \mathbf{R}_+, \\ u(t) = g(u)(t), & t \in [\,-\tau, 0\,], \end{cases} \tag{15}$$

where $A_\varepsilon = A - \varepsilon I$ and $f_\varepsilon(t) = f(t) + \varepsilon z(t)$ for $t \in \mathbf{R}_+$. Clearly, $A_\varepsilon + \varepsilon I$ is m-dissipative, $0 \in D(A_\varepsilon)$ and $0 \in A_\varepsilon 0$. Since $z \in C_b([\,0, +\infty); \overline{D(A)})$ we have $f_\varepsilon \in L^\infty(\mathbf{R}_+; X) \cap L^1(\mathbf{R}_+; X)$ and so Lemma 1 applies with $\omega = \varepsilon$ and this implies the existence and uniqueness of solution $u_\varepsilon^f \in C_b([\,-\tau, +\infty); \overline{D(A)})$.

Next, using the very same operator $A_\varepsilon = A - \varepsilon I$, we rewrite the unperturbed problem (9) as

$$\begin{cases} z'(t) \in A_\varepsilon z(t) + h_\varepsilon(t), & t \in \mathbf{R}_+, \\ z(t) = g(z)(t), & t \in [\,-\tau, 0\,], \end{cases} \tag{16}$$

with $h_\varepsilon(t) = \varepsilon z(t)$, for $t \in \mathbf{R}_+$. Then, for each $t \in (0, +\infty)$, the unique C^0-solution u_ε^f of (15) and the unique solution z of (16) satisfy

$$\|u_\varepsilon^f(t) - z(t)\| \leq e^{-\varepsilon t}\|u_\varepsilon^f(0) - z(0)\| + \int_0^t e^{-\varepsilon(t-s)}\|f(s)\|\,ds$$

$$\leq e^{-\varepsilon t}\|u_\varepsilon^f - z\|_{C_b([\,a,+\infty);X)} + \frac{1 - e^{-\varepsilon t}}{\varepsilon}\|f\|_{L^\infty(\mathbf{R}_+;X)},$$

for each $t \in (0, +\infty)$.

Clearly, there exists a sequence (α_n) in $(0, a)$ such that

$$\lim_n \|u_\varepsilon^f - z\|_{C_b([\,\alpha_n,+\infty);X)} = \|u_\varepsilon^f - z\|_{C_b([0,+\infty);X)}. \tag{17}$$

From the last inequality it follows that, for every $n \in \mathbf{N}$, we have

$$\|u_\varepsilon^f(t) - z(t)\| \le e^{-\varepsilon\alpha_n}\|u_\varepsilon^f - z\|_{C_b([\,\alpha_n,+\infty);X)} + \frac{1-e^{-\varepsilon\alpha_n}}{\varepsilon}\|f\|_{L^\infty(\mathbf{R}_+;X)} \quad (18)$$

for each $t \in [\alpha_n, +\infty)$, and so

$$\|u_\varepsilon^f - z\|_{C_b([\,\alpha_n,+\infty);X)} \le \frac{1}{\varepsilon}\|f\|_{L^\infty(\mathbf{R}_+;X)},$$

for every $n \in \mathbf{N}$. From (17), it readily follows that

$$\|u_\varepsilon^f - z\|_{C_b([\,0,+\infty);X)} \le \frac{1}{\varepsilon}\|f\|_{L^\infty(\mathbf{R}_+;X)}.$$

Next, if $t \in [\,-\tau, 0\,]$, from (g_2) in (H_g), we get

$$\|u_\varepsilon^f(t) - z(t)\| = \|g(u_\varepsilon^f)(t) - g(z)(t)\|$$

$$\le \|u_\varepsilon^f - z\|_{C_b([\,a,+\infty);X)} \le \|u_\varepsilon^f - z\|_{C_b([\,0,+\infty);X)}$$

and thus (14) holds true, and this completes the proof. □

Lemma 3 *Let us assume that (A_1), (A_2) in (H_A) and (H_g) are satisfied, let $\varepsilon > 0$ be fixed and let $\ell \in L^\infty(\mathbf{R}_+;\mathbf{R}_+) \cap L^1(\mathbf{R}_+;\mathbf{R}_+)$. Then the operator $f \mapsto u_\varepsilon^f$, where u_ε^f is the unique solution of the problem* (11) *corresponding to f, maps the set*

$$\mathcal{F} = \{f \in L^\infty([\,0,+\infty);X) \cap L^1(\mathbf{R}_+;X);\ \|f(t)\| \le \ell(t) \ \ \textit{a.e. for } t \in \mathbf{R}_+\},$$

into a relatively compact set in $\widetilde{C}_b([\,-\tau,+\infty);\overline{D(A)})$.

Proof. By (14), $\{u_\varepsilon^f;\ f \in \mathcal{F}\}$ is bounded in $C_b([\,0,+\infty);\overline{D(A)})$ and thus $\{u_\varepsilon^f(0);\ f \in \mathcal{F}\}$ is bounded in $\overline{D(A)}$. Since $\mathcal{F}$ is uniformly integrable in $L^1(0,k;X)$ for $k = 1,2,\ldots$ – see Definition 3 –, from (A_2) and Theorem 5, we conclude that, for every $k = 1,2,\ldots$, and $\sigma \in (0,k)$, $\{u_\varepsilon^f; f \in \mathcal{F}\}$ is relatively compact in $C([\,\sigma,k\,];\overline{D(A)})$. Thanks to (g_2), (g_4) in (H_g) and to Remark 6, we deduce that the set $\{g(u_\varepsilon^f);\ f \in \mathcal{F}\}$ is relatively compact in $C([\,-\tau,0\,];\overline{D(A)})$, and therefore $\Big\{g(u_\varepsilon^f)(0);\ f \in \mathcal{F}\Big\} = \{u_\varepsilon^f(0);\ f \in \mathcal{F}\}$ is relatively compact in $\overline{D(A)}$. Again, from (g_1) and the second part of Theorem 5, it follows that the set $\{u_\varepsilon^f;\ f \in \mathcal{F})\}$ is relatively compact in $\widetilde{C}_b([\,-\tau,+\infty);\overline{D(A)})$. The proof is complete. □

Lemma 4 *Let us assume that* (H_A), (H_F), (H'_B) *and* (H_g) *are satisfied. Then, for each* $\varepsilon > 0$, *the problem* (13) *has at least one solution* u_ε.

Since the proof follows the very same lines as those in the proof of Lemma 4.3 in Vrabie [47], we do not give details.

□

Lemma 5 *If* (H_A), (H_F), (H'_I), (H'_B) *and* (H_g) *are satisfied, then, for each* $\varepsilon \in (0,1)$, *each* C^0*-solution* u_ε *of the problem* (13) *satisfies*

$$\|u_\varepsilon - z\|_{C_b([0,+\infty);X)} \leq r, \tag{19}$$

where $r > 0$ *is given by* (H'_I).

Proof. Let us observe that, if $0 \leq t < \widetilde{t}$, we have

$$\|u_\varepsilon(\widetilde{t}) - z(\widetilde{t})\| \leq \|u_\varepsilon(t) - z(t)\| \tag{20}$$

$$+ \int_t^{\widetilde{t}} [\, u_\varepsilon(s) - z(s), f(s)\,]_+ \, ds - \varepsilon \int_t^{\widetilde{t}} \|u_\varepsilon(s) - z(s)\| \, ds.$$

Let us assume by contradiction that there exists $t \in \mathbf{R}_+$ such that

$$\|u_\varepsilon(t) - z(t)\| > r.$$

We distinguish between two cases.

Case 1. There exists $t_m \in \mathbf{R}_+$ such that

$$r < \|u_\varepsilon - z\|_{C_b([0,+\infty);X)} = \|u_\varepsilon(t_m) - z(t_m)\|. \tag{21}$$

If $t_m = 0$, then

$$r < \|u_\varepsilon - z\|_{C_b([0,+\infty);X)} = \|u_\varepsilon(0) - z(0)\| = \|g(u_\varepsilon)(0) - g(z)(0)\|$$

$$\leq \|u_\varepsilon - z\|_{C_b([a,+\infty);X)} \leq \|u_\varepsilon - z\|_{C_b([0,+\infty);X)}$$

and so

$$\|u_\varepsilon - z\|_{C_b([0,+\infty);X)} = \|u_\varepsilon - z\|_{C_b([a,+\infty);X)}.$$

Therefore, we can always confine ourselves to analyze the case when, in (21), either $t_m \in (0,+\infty)$ or there is no $t_m \in (0,+\infty)$ satisfying the equality in (21).

So, if there exists $t_m \in (0,+\infty)$ such that (21) holds true, then the mapping

$$t \mapsto \|u_\varepsilon(t) - z(t)\|$$

cannot be constant on $(0,t_m)$. Indeed, if we assume that

$$\|u_\varepsilon(s) - z(s)\| = \|u_\varepsilon(t_m) - z(t_m)\|$$

for each $s \in (0,t_m)$, then, taking $t \in (0,t_m)$ and $\widetilde{t} = t_m$ in (20) and using (H_I') with $v(0) = u_{\varepsilon s}(0) = u_\varepsilon(s)$, we get

$$r < r - \varepsilon(t_m - t)r < r \tag{22}$$

which is impossible. Consequently, there exists $t_0 \in (0,t_m)$ such that

$$r < \|u_\varepsilon(t_0) - z(t_0)\| < \|u_\varepsilon(s) - z(s)\| \leq \|u_\varepsilon(t_m) - z(t_m)\| = \|u_\varepsilon - z\|_{C_b([\,0,+\infty);X)}$$

for each $s \in (t_0,t_m)$. Since

$$\|u_\varepsilon(s) - z(s)\| \leq \|u_{\varepsilon s} - z_s\|_{C([\,-\tau,0\,];X)},$$

for each $s \in \mathbf{R}_+$, we have

$$r < \|u_{\varepsilon s} - z_s\|_{C([\,-\tau,0\,];X)}$$

for each $s \in (t_0,t_m)$ and then, using again (20) and (H_I'), we get

$$r < \|u_\varepsilon(t_m) - z(t_m)\| \leq \|u_\varepsilon(t_0) - z(t_0)\| - \varepsilon(t_m - t_0)r$$

which implies the very same contradiction as before, i.e. (22).

It remains only to analyze

Case 2. There is no $t_m \in \mathbf{R}_+$ such that (21) holds true. Then, there exists at least one sequence $(t_k)_k$ such that

$$\begin{cases} \lim\limits_k t_k = +\infty, \\ \lim\limits_k \|u_\varepsilon(t_k) - z(t_k)\| = \|u_\varepsilon - z\|_{C_b([\,0,+\infty);X)}. \end{cases}$$

If there exists $\widetilde{t} \in \mathbf{R}_+$ such that $\|u_\varepsilon(\widetilde{t}) - z(\widetilde{t})\| = r$, then $\|u_\varepsilon(t) - z(t)\| \leq r$ for each $t \in [\,\widetilde{t},+\infty)$. Indeed, if we assume the contrary, there would exists $[\,t,\widetilde{t}\,] \subseteq [\,0,+\infty)$ such that

$$\|u_\varepsilon(t) - z(t)\| = r$$

and

$$r < \|u_\varepsilon(s) - z(s)\|$$

for each $s \in (t, \widetilde{t}\,]$. Then, using once again (20) and (H_I'), we get

$$r < \|u_\varepsilon(\widetilde{t}) - z(\widetilde{t})\| \le \|u_\varepsilon(t) - z(t)\| - \varepsilon(\widetilde{t} - t)r$$

$$\le r - \varepsilon(\widetilde{t} - t)r$$

leading to (22) which is impossible.

So, when both

$$r < \|u_\varepsilon - z\|_{C_b([0,+\infty);X)}$$

and

$$\|u_\varepsilon(t) - z(t)\| < \|u_\varepsilon - z\|_{C_b([0,+\infty);X)}$$

hold true for each $t \in \mathbf{R}_+$, we necessarily have

$$\|u_\varepsilon(t) - z(t)\| > r$$

for each $t \in \mathbf{R}_+$. If this is the case, let us remark that we may assume with no loss of generality, by extracting a subsequence if necessary, that

$$t_{k+1} - t_k \ge 1$$

for $k = 0, 1, 2, \ldots$. Then, by (3) and (H_I'), we have

$$r < \|u_\varepsilon(t_{k+1}) - z(t_{k+1})\|$$

$$\le \|u_\varepsilon(t_k) - z(t_k)\| + \int_{t_k}^{t_{k+1}} [\, u_\varepsilon(s) - z(s), f(s) - \varepsilon(u_\varepsilon(s) - z(s))\,]_+ \, ds$$

$$\le \|u_\varepsilon(t_k) - z(t_k)\| - \varepsilon \int_{t_k}^{t_{k+1}} \|u_\varepsilon(s) - z(s)\| \, ds$$

$$\le \|u_\varepsilon(t_k) - z(t_k)\| - \varepsilon(t_{k+1} - t_k)r \le \|u_\varepsilon(t_k) - z(t_k)\| - \varepsilon r$$

for each $k \in \mathbf{N}$. Passing to the limit for $k \to +\infty$ in the inequalities

$$\|u_\varepsilon(t_{k+1}) - z(t_{k+1})\| \le \|u_\varepsilon(t_k) - z(t_k)\| - \varepsilon r, \ k = 1, 2, \ldots$$

we get

$$\|u_\varepsilon - z\|_{C_b([0,+\infty);X)} \le \|u_\varepsilon - z\|_{C_b([0,+\infty);X)} - \varepsilon r.$$

But, in view of Lemma 2, $\|u_\varepsilon - z\|_{C_b([0,+\infty);X)}$ is finite and thus we get a contradiction. This contradiction can be eliminated only if **Case 2** cannot hold. Thus, both **Case 1** and **Case 2** are impossible. In turn, this is a

contradiction too, because at least one of these two cases should hold true. So, the initial supposition, that $\|u_\varepsilon - z\|_{C_b([\,0,+\infty);X)} > r$, is necessarily false. It then follows that (19) holds true and this completes the proof. □

Now, we can pass to the proof of Theorem 8.

Proof. Let $(\varepsilon_n)_n$ be a sequence with $\varepsilon_n \downarrow 0$, let $(u_n)_n$ be the sequence of the C^0-solutions of the problem (13) corresponding to $\varepsilon = \varepsilon_n$ for $n \in \mathbf{N}$, and let $(f_n)_n$ be such that

$$\begin{cases} u_n'(t) \in Au_n(t) - \varepsilon_n[\,u_n(t) - z(t)\,] + f_n(t), & t \in \mathbf{R}_+, \\ f_n(t) \in F_{\varepsilon_n}(t, u_{nt}), & t \in \mathbf{R}_+, \\ u_n(t) = g(u_n)(t), & t \in [\,-\tau, 0\,]. \end{cases}$$

In view of Remark 1, we may assume without loss of generality that $E_{\varepsilon_{n+1}} \subset E_{\varepsilon_n}$ for $n = 0, 1, \ldots$. This means that

$$F_{\varepsilon_n}(t, v) = F_{\varepsilon_{n+1}}(t, v) \tag{23}$$

for each $t \in \mathbf{R}_+ \setminus E_{\varepsilon_n}$ and $v \in C([\,-\tau, 0\,]; \overline{D(A)})$.

From (H_B'), we deduce that, for $k = 1, 2, \ldots$, the set $\{f_n;\ n \in \mathbf{N}\}$ is uniformly integrable in $L^1(0, k; X)$. Then, from Lemma 5, (A_2) in (H_A) and Theorem 5, it follows that, for $k = 1, 2, \ldots$, and each $\sigma \in (0, k)$, the set $\{u_n;\ n \in \mathbf{N}\}$ is relatively compact in $C([\,\sigma, k\,]; \overline{D(A)})$. In view of (g_4) in (H_g), we deduce that the set $\{u_n;\ n \in \mathbf{N}\}$ is relatively compact in $C([\,-\tau, 0\,]; \overline{D(A)})$. In particular, the set

$$\{u_n(0) = g(u_n)(0);\quad n \in \mathbf{N}\}$$

is relatively compact in $\overline{D(A)}$. From the second part of Theorem 5, we conclude that $\{u_n;\ n \in \mathbf{N}\}$ is relatively compact in $C([\,0, k\,]; \overline{D(A)})$ for $k = 1, 2, \ldots$ and thus in $C([\,-\tau, k\,]; \overline{D(A)})$. So, $\{u_n;\ n \in \mathbf{N}\}$ is relatively compact in $\widetilde{C}_b([\,-\tau, +\infty); \overline{D(A)})$. Accordingly, for each $k = 1, 2, \ldots$,

$$C_k = \overline{\{u_n(t);\quad n \in \mathbf{N},\quad t \in [\,0, k\,]\}}$$

is compact in $\overline{D(A)}$. Let $\gamma \in (0, 1)$ be arbitrary, let E_γ be the Lebesgue measurable set in $[\,0, +\infty)$ given by Definition 2 and, for each $k = 1, 2, \ldots$, let us define the set

$$D_{\gamma,k} = \overline{\bigcup_{n \in \mathbf{N}} \{(t, u_{\varepsilon_n t});\quad t \in [\,0, k\,] \setminus E_\gamma\}}.$$

Clearly, $D_{\gamma,k}$ is compact in $\mathbf{R}_+ \times C([-\tau, 0]; \overline{D(A)})$. Next, for each $\gamma \in (0,1)$ and each $k = 1, 2, \dots$, let us define

$$C_{\gamma,k} = F_\gamma(D_{\gamma,k}) = F(D_{\gamma,k}) \cup \{0\}$$

which is weakly compact since $D_{\gamma,k}$ is compact and $F_{|D_{\gamma,k}}$ is strongly-weakly u.s.c. See Lemma 2.6.1, p. 47 in Cârjă, Necula, Vrabie [19]. Further, the family $\mathcal{F} = \{f_{\varepsilon_n};\ n = 0, 1, \dots\} \subseteq L^1(\mathbf{R}_+; X)$ satisfies the hypotheses of Theorem 4.1 in Vrabie [46]. So, on a subsequence at least, we have

$$\begin{cases} \lim\limits_n f_n = f \quad \text{weakly in } \ L^1(\mathbf{R}_+; X), \\ \lim\limits_n u_n = u \quad \text{in } \ \widetilde{C}_b([-\tau, +\infty); \overline{D(A)}), \\ \lim\limits_n u_{nt} = u_t \ \text{ in } \ C([-\tau, 0]; \overline{D(A)}) \ \text{ for each } t \in \mathbf{R}_+. \end{cases}$$

From Lemma 2.6.2, p. 47 in Cârjă, Necula, Vrabie [19] combined with (23), we get

$$f(t) \in F_{\varepsilon_n}(t, u_t)$$

for each $n \in \mathbf{R}$ and a.e. $t \in \mathbf{R}_+ \setminus E_{\varepsilon_n}$. Since $\lim_n \lambda(E_{\varepsilon_n}) = 0$, it follows that

$$f(t) \in F(t, u_t)$$

a.e. $t \in \mathbf{R}_+$. But A is of complete continuous type, wherefrom it follows that u is a C^0-solution of the problem (1) corresponding to the selection f of $t \mapsto F(t, u_t)$. Finally, it suffices to observe that, from (19) in Lemma 5, it follows that $u(t) - z(t) \in D(0, r)$ for each $t \in \mathbf{R}_+$. □

We can now proceed to the proof of Theorem 7.

Proof. Let $r > 0$ be given by (H_I) and let us define the set

$$\mathcal{K}_r = \{(t, v) \in \mathbf{R}_+ \times C([-\tau, 0]; \overline{D(A)});\ \|v(0) - z(t)\| \le r\}.$$

Clearly, $\mathcal{K}_r$ is nonempty and closed in $\mathbf{R}_+ \times C([-\tau, 0]; X)$, In addition, since by (A_1) in (H_A), $\overline{D(A)}$ is convex, it follows that for each $t \in \mathbf{R}_+$, the cross-section of $\mathcal{K}_r$ at t, i.e.

$$\mathcal{K}_r(t) = \{v \in C([-\tau, 0]; \overline{D(A)});\ (t, v) \in \mathcal{K}_r\}$$

is convex. Let $\pi : \mathbf{R}_+ \times C([-\tau, 0]; \overline{D(A)}) \to \mathbf{R}_+ \times C([-\tau, 0]; X)$ be defined by

$$\pi(t, v) = \begin{cases} (t, v) & \text{if } \ \|v(0) - z(t)\| \le r, \\ \left(t, \dfrac{r}{\|v - z_t\|_{C([-\tau,0];X)}}(v - z_t) + z_t\right) & \text{if } \ \|v(0) - z(t)\| > r. \end{cases}$$

We observe that π is continuous, π restricted to $\mathcal{K}_r$ is the identity operator and π maps $\mathbf{R}_+ \times C([-\tau, 0]; \overline{D(A)})$ into $\mathcal{K}_r$. The first two properties mentioned are obvious. To prove the fact that π maps $\mathbf{R}_+ \times C([-\tau, 0]; \overline{D(A)})$ into $\mathcal{K}_r$, we have merely to observe that if $\|v(0) - z(t)\| > r$ then, inasmuch as $\overline{D(A)}$ is convex and $v, z_t \in C([-\tau, 0]; \overline{D(A)}$, it follows that their convex combination

$$\frac{r}{\|v - z_t\|_{C([-\tau,0];X)}}(v - z_t) + z_t - z_t \in C([-\tau, 0]; \overline{D(A)}).$$

Moreover

$$\left\| \frac{r}{\|v - z_t\|_{C([-\tau,0];X)}}(v - z_t) + z_t - z_t \right\|_{C([-\tau,0];X)} = r$$

and so, in this case, $\pi(t, v) \in \mathcal{K}_r$. If $\|v(0) - z(t)\| \leq r$, then $\pi(t, v) = (t, v)$ and thus, π maps $\mathbf{R}_+ \times C([-\tau, 0]; \overline{D(A)})$ into $\mathcal{K}_r$.

Then, we can define the multifunction $F_\pi : \mathbf{R}_+ \times C([-\tau, 0]; \overline{D(A)}) \hookrightarrow X$ by

$$F_\pi(t, v) = F(\pi(t, v)),$$

for each $(t, v) \in \mathbf{R}_+ \times C([-\tau, 0]; \overline{D(A)})$. As π is continuous, it follows that F_π satisfies (H_F). Moreover, one can easily verify that it satisfies (H'_B). Moreover, since

$$\pi(\mathbf{R}_+ \times C([-\tau, 0]; \overline{D(A)})) \subseteq \mathcal{K}_r,$$

we conclude that F_π satisfies (H'_I) too. Indeed, let $(t, v) \in \mathbf{R}_+ \times C([-\tau, 0]; \overline{D(A)})$ be arbitrary and satisfying

$$\|v(0) - z(t)\| > r \tag{24}$$

and let $f \in F(\pi(t, v))$.

From the definition of π, it follows that the projection P_2 of $\pi(t, v)$ on the second component, i.e.

$$P_2(\pi(t, v)) = \begin{cases} v & \text{if } \|v(0) - z(t)\| \leq r, \\ \dfrac{r}{\|v - z_t\|_{C([-\tau,0];X)}}(v - z_t) + z_t & \text{if } \|v(0) - z(t)\| > r. \end{cases}$$

satisfies:

$$\|P_2(\pi(t, v)) - z_t\|_{C([-\tau,0];X)} = \begin{cases} r & \text{if } \|v(0) - z(t)\| > r, \\ \|v - z_t\|_{C([-\tau,0];X)} & \text{if } \|v(0) - z(t)\| \leq r. \end{cases}$$

Therefore, if (t, v) satisfies (24), it follows that

$$\|P_2(\pi(t,v)) - z_t\|_{C([-\tau,0];X)} = r.$$

So, by (H_I), we have

$$[\, v(0) - z(t), f \,]_+ = [\, P_2(\pi(t,v))(0) - z(t), f \,]_+ \leq 0$$

which proves that F_π satisfies (H_I').

Hence, by virtue of Theorem 8, the problem

$$\begin{cases} u'(t) = Au(t) + f(t), & t \in \mathbf{R}_+, \\ f(t) \in F_\pi(t, u_t), & t \in \mathbf{R}_+, \\ u(t) = g(u)(t), & t \in [\,-\tau, 0\,] \end{cases}$$

has at least one C^0-solution $u \in C_b([\,-\tau, +\infty); \overline{D(A)})$.

By (19), we have $\|u_t(0) - z(t)\| \leq r$ for each $t \in \mathbf{R}_+$. So, $(t, u_t) \in \mathcal{K}_r$, which shows that

$$F_\pi(t, u_t) = F(t, u_t)$$

for each $t \in \mathbf{R}_+$. Thus u is a C^0-solution of (1) and this completes the proof of Theorem 7. □

7 Nonlinear diffusion in $L^1(\Omega)$

Let Ω be a nonempty, bounded and open subset in $\mathbf{R}^d$, $d \geq 1$, with C^1 boundary Σ, let $\varphi : D(\varphi) \subseteq \mathbf{R} \hookrightarrow \mathbf{R}$ be maximal monotone with $0 \in \varphi(0)$ and let $\omega > 0$. Let us consider the porous medium equation subjected to nonlocal initial conditions

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) \in \Delta\varphi(u(t,x)) - \omega u(t,x) + f(t,x), & \text{in } Q_+, \\ f(t,x) \in F\left(t, u(t), \displaystyle\int_{-\tau}^{0} u(t+s,x)\, ds\right), & \text{in } Q_+, \\ \varphi(u(t,x)) = 0, & \text{on } \Sigma_+, \\ u(t,x) = \displaystyle\int_{\tau}^{+\infty} \mathcal{N}(u(\theta+t))(x)\, d\mu(\theta) + \psi(t)(x), & \text{in } Q_\tau. \end{cases} \tag{25}$$

Let us consider the auxiliary problem

$$\begin{cases} \dfrac{\partial z}{\partial t}(t,x) \in \Delta\varphi(z(t,x)) - \omega z(t,x), & \text{in } Q_+, \\ \varphi(z(t,x)) = 0, & \text{on } \Sigma_+, \\ z(t,x) = \displaystyle\int_\tau^{+\infty} \mathcal{N}(z(\theta+t))(x)\, d\mu(\theta) + \psi(t)(x), & \text{in } Q_\tau \end{cases} \tag{26}$$

and let us denote by $z \in C_b([\,-\tau,+\infty); L^1(\Omega))$ the unique C^0-solution of (26).

Before passing to the statement of the main existence result concerning (25), we need to introduce some notation and to explain the exact definition of F.

Let $f_i : \mathbf{R}_+ \times \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ be two functions with $f_1(t,u,v) \le f_2(t,u,v)$ for each $(t,u,v) \in \mathbf{R}_+ \times \mathbf{R} \times \mathbf{R}$ and let

$$F : \mathbf{R}_+ \times C([\,-\tau,0\,]; L^1(\Omega)) \hookrightarrow L^1(\Omega)$$

be given by

$$F := F_0 + F_1,$$

where

$$F_0(t,v) = \left\{ f \in L^1(\Omega); f(x) \in [\,\widetilde{f}_1(t,v)(x), \widetilde{f}_1(t,v)(x)\,], \ \text{ a.e. for } x \in \Omega \right\}$$

and

$$F_1(t,v)(x) := \{\sigma(t)h(x)\}$$

for each $(t,v) \in \mathbf{R}_+ \times C([\,-\tau,0\,]; L^1(\Omega))$. Here

$$\widetilde{f}_i : \mathbf{R}_+ \times \Omega \times C([\,-\tau,0\,]; L^1(\Omega)) \to \mathbf{R}, \ i = 1,2,$$

are defined as:

$$\begin{cases} \widetilde{f}_1(t,x,v) := f_1\left(t, v(0)(x), \displaystyle\int_{-\tau}^0 v(s)(x)\, ds\right) \\ \widetilde{f}_2(t,x,v) := f_2\left(t, v(0)(x), \displaystyle\int_{-\tau}^0 v(s)(x)\, ds\right) \end{cases} \tag{27}$$

for each $(t,v) \in \mathbf{R}_+ \times C([\,-\tau,0\,]; L^1(\Omega))$, a.e. in Ω, $h \in L^1(\Omega)$ is a fixed element satisfying $\|h\|_{L^1(\Omega)} \neq 0$ and $\sigma \in L^1(\mathbf{R}_+; \mathbf{R})$.

Theorem 9 *Let Ω be a nonempty, bounded and open subset in $\mathbf{R}^d$ with C^1 boundary Σ, let $\omega > 0$ and let $\varphi : \mathbf{R} \to \mathbf{R}$ be continuous on $\mathbf{R}$ and C^1 on $\mathbf{R} \setminus \{0\}$ with $\varphi(0) = 0$ and for which there exist two constants $C > 0$ and $\alpha > 0$ if $d \leq 2$ and $\alpha > (d-2)/d$ if $d \geq 3$ such that*

$$\varphi'(r) \geq C|r|^{\alpha-1}$$

for each $r \in \mathbf{R} \setminus \{0\}$. Let $f_i : \mathbf{R}_+ \times \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ be two given functions, $h \in L^1(\Omega)$, $\|h\|_{L^1(\Omega)} > 0$, $\sigma \in L^1(\mathbf{R}_+; \mathbf{R})$ and let F be defined as above.

Let $\mathcal{N} : L^1(\Omega) \to L^1(\Omega)$, $\psi \in C([-\tau, 0]; L^1(\Omega))$ and let μ be a σ-finite and complete measure on $[\tau, +\infty)$. Let us assume that:

(σ_1) *$\|\sigma(t)\| \leq 1$ for each $t \in \mathbf{R}_+$;*

(F_1) *$f_1(t, u, v) \leq f_2(t, u, v)$ for each $(t, u, v) \in \mathbf{R}_+ \times \mathbf{R} \times \mathbf{R}$;*

(F_2) *f_1 is l.s.c. and f_2 is u.s.c. and, for each $(t, u, v), (t, u, w) \in \mathbf{R}_+ \times \mathbf{R} \times \mathbf{R}$ with $v \leq w$, we have*

$$\begin{cases} f_1(t, u, v) \leq f_1(t, u, w), \\ f_2(t, u, v) \geq f_2(t, u, w); \end{cases}$$

(F_3) *there exists $c > 0$ such that, for every $(t, x, v) \in D(f_1, f_2)$ with*

$$\|v(0)(\cdot) - z(t, \cdot)\|_{L^1(\Omega)} \leq c^{-1}\|h\|_{L^1(\Omega)}$$

we have

$$sign\,[\,v(0)(x) - z(t, x)\,]f_0(x) \leq -c|v(0)(x) - z(t, x)|$$

for each $f_0(x) \in [\,f_1(t, x, v), f_2(t, x, v)\,]$, z being the unique C^0-solution of the problem (26);

(F_4) *there exists a nonnegative function $\widetilde{\ell} \in L^1(\mathbf{R}_+; \mathbf{R}) \cap L^\infty(\mathbf{R}_+; \mathbf{R})$ such that*

$$|f_i(t, u, v)| \leq \widetilde{\ell}(t)$$

for $i = 1, 2$ and for each $(t, u, v) \in \mathbf{R}_+ \times \mathbf{R} \times \mathbf{R}$;

(F_5) *for each $t \in \mathbf{R}_+$ and each $v \in C([-\tau, 0]; L^1(\Omega))$, we have*

$$f_i(t, z(t, x), v) = 0$$

for $i = 1, 2$ and a.e. for $x \in \Omega$;

(μ_1) *there exists* $b > \tau$ *such that* $\operatorname{supp}\mu \subseteq [\,b, +\infty)$;

(μ_2) $\mu([\,b, \infty)) = 1$;

($\mathcal{N}_1$) $\|\mathcal{N}(u) - \mathcal{N}(v)\|_{L^1(\Omega)} \leq \|u - v\|_{L^1(\Omega)}$ *for each* $u, v \in L^1(\Omega)$;

($\mathcal{N}_2$) $\mathcal{N}(0) = 0$.

Then, the problem (25) *has at least one* C^0*-solution* $u \in C_b([\,-\tau, +\infty); L^1(\Omega))$ *satisfying*

$$\|u - z\|_{C_b(\mathbf{R}_+; L^1(\Omega))} \leq c^{-1}\|h\|_{L^1(\Omega)}.$$

Remark 7 Condition (F_5) is satisfied, for instance, if

$$f_i(t, u, v) = \psi(t, u) \cdot \overline{f}_i(t, u, v),$$

where ψ is positive, continuous and bounded and $\psi(t, z(t, x)) = 0$, while $\overline{f}_i$ satisfy (F_1) $\sim$ (F_4), $i = 1, 2$. In the particular case in which $\psi \equiv 0$, it follows that $z \equiv 0$ and so, (F_5) reduces to

$$f_i(t, 0, v) = 0$$

for each $(t, v) \in \mathbf{R}_+ \times \mathbf{R}$.

Proof. Let $X = L^1(\Omega)$ and let us define $A : D(A) \subseteq L^1(\Omega) \to L^1(\Omega)$, by

$$Au := \Delta\varphi(u) - \omega u$$

for each $u \in D(A)$, where

$$D(A) = \left\{u \in L^1(\Omega);\ \varphi(u) \in W_0^{1,1}(\Omega),\ \Delta\varphi(u) \in L^1(\Omega)\right\}.$$

As $\varphi(0) = 0$, $C_0^\infty(\Omega)$ is dense in $D(A)$ and so $\overline{D(A)} = L^1(\Omega)$.

Theorem 6 implies that A is m-dissipative and $A + \omega I$ is dissipative in $L^1(\Omega)$, $A0 = 0$, A generates a compact semigroup and is of complete continuous type on $\overline{D(A)} = L^1(\Omega)$. Hence, A satisfies (H_A). Let F be defined as above and

$$g : C_b([\,-\tau, +\infty); L^1(\Omega)) \to C([\,-\tau, 0\,]; L^1(\Omega))$$

be defined by

$$g(u)(t)(x) = \int_\tau^{+\infty} \mathcal{N}(u(t + \theta))(x)\, d\mu(\theta) + \psi(t)(x)$$

for each $u \in C_b([\,-\tau,+\infty);L^1(\Omega))$, each $t \in [\,-\tau,0\,]$ and a.e. for $x \in \Omega$.

From (σ_1), (F_1), (F_2), (F_4) and Lemma 5.1 in Vrabie [47], using a similar arguments as in the proof of the corresponding part in the preceding section, we conclude that F satisfies (H_F). From (F_2) and (F_3), we conclude that F satisfies (H_I) and (H_B) with

$$r = c^{-1}\|h\|_{L^1(\Omega)}.$$

Indeed, we will show that for each $(t,v) \in \mathbf{R}_+ \times C([\,-\tau,0\,];L^1(\Omega))$, with

$$\|v(0)(\cdot) - z(t,\cdot)\|_{L^1(\Omega)} = r,$$

and every $f \in F(t,v)$, we have

$$[\,v(0)(\cdot) - z(t,\cdot), f\,]_+ \le 0.$$

Let us observe that in our case, i.e. $X = L^1(\Omega)$, we have

$$[\,v(0)(\cdot) - z(t,\cdot), f\,]_+ = \int_{\{y\in\Omega;v(0)(y)-z(t,y)>0\}} f(x)\,dx$$

$$-\int_{\{y\in\Omega;v(0)(y)-z(t,y)<0\}} f(x)\,dx + \int_{\{y\in\Omega;v(0)(y)-z(t,y)=0\}} |f(x)|\,dx.$$

Let $f \in F(t,v)$. Clearly f is of the form $f = f_0 + h$, where $f_0 \in L^1(\Omega)$ satisfies $f_1(t,x,v) \le f_0(x) \le f_2(t,x,v)$ a.e. for $x \in \Omega$. From the definition of $[\,\cdot,\cdot\,]_+$ in $L^1(\Omega)$, we deduce

$$[\,v(0)(\cdot) - z(t,\cdot), f\,]_+$$

$$\le \int_{\{y\in\Omega;v(0)(y)-z(t,y)>0\}} f_0(x)\,dx - \int_{\{y\in\Omega;v(0)(y)-z(t,y)<0\}} f_0(x)\,dx$$

$$+\int_{\{y\in\Omega;v(0)(y)-z(t,y)=0\}} |f_0(x)|\,dx + \int_{\{y\in\Omega;v(0)(y)-z(t,y)>0\}} \alpha(t)h(x)\,dx$$

$$-\int_{\{y\in\Omega;v(0)(y)-z(t,y)<0\}} h(x)\,dx + \int_{\{y\in\Omega;v(0)(y)-z(t,y)=0\}} |\alpha(t)|\cdot|h(x)|\,dx.$$

Next, taking into account that, from (F_5), we have $f_0(x) = 0$ a.e. for those $x \in \Omega$ for which $v(0)(x) = z(t,x)$, the last inequality, conjunction with (F_4), yields

$$[\,v(0)(\cdot)-z(t,\cdot), f\,]_+ \le \int_\Omega \operatorname{sign}[\,v(0)(x)-z(t,x)\,]f_0(x)\,dx + \int_\Omega |\alpha(t)|\cdot|h(x)|\,dx$$

$$\leq -c\int_\Omega |v(0)(x)-z(t,x)|\,dx+\int_\Omega |h(x)|\,dx\leq 0.$$

So, F satisfies (H_I). As (H_4) follows from (F_3), we deduce that F satisfies (H_B). Since the proof of (H_g) is very simple, we do not enter into details. So, we are in the hypotheses of Theorem 7 wherefrom the conclusion.

Acknowledgments

The work of both authors was supported by Grant PN-II-ID-PCE-2011-3-0052 of CNCS Romania.

References

[1] S. Aizicovici, H. Lee. Nonlinear nonlocal Cauchy problems in Banach spaces. *Appl. Math. Lett.* 18:401-407, 2005.

[2] S. Aizicovici, M. McKibben. Existence results for a class of abstract nonlocal Cauchy problems. *Nonlinear Anal.* 39: 649–668, 2000.

[3] S. Aizicovici, N. S. Papageorgiou, V. Staicu. Periodic solutions of nonlinear evolution inclusions in Banach spaces. *J. Nonlinear Convex Anal.* 7:163-177, 2006.

[4] S. Aizicovici, N. H. Pavel, I. I. Vrabie. Anti-periodic solutions to strongly nonlinear evolution equations in Hilbert spaces. *An. Ştiinţ. Univ. Al. I. Cuza Iaşi (N.S.), Secţ. I a Mat.* 44:227-234, 1998.

[5] S. Aizicovici, V. Staicu. Multivalued evolution equations with nonlocal initial conditions in Banach spaces. *NoDEA Nonlinear Differential Equations Appl.* 14:361-376, 2007.

[6] J. Andres. Periodic-type solutions of differential inclusions. *Adv. Math. Res.* 8:295-353, 2009.

[7] O. Arino, S. Gautier, J. P. Penot. A Fixed Point theorem for sequentially continuous mappings with applications to ordinary differential equations. *Funkcial. Ekvac.* 27:273-279, 1984.

[8] G. Avalishvili, M. Avalishvili. Nonclassical problems with nonlocal initial conditions for abstract second-order evolution equations. *Bull. Georgian Natl. Acad. Sci. (N.S.)* 5:17-24, 2011.

[9] M. Badii, J. I. Díaz, A. Tesei. Existence and attractivity results for a class of degenerate functional parabolic problems. *Rend. Semin. Mat. Univ. Padova* 78:109-124, 1987.

[10] P. Baras. Compacité de l'opérateur definissant la solution d'une équation d'évolution non linéaire $(du/dt) + Au \ni f$. *C. R. Math. Acad. Sci. Paris* 286:1113-1116, 1978.

[11] V. Barbu. *Nonlinear Differential Equations of Monotone Type in Banach Spaces*. Springer Monographs in Mathematics, Berlin, 2010.

[12] P. Benilan. *Equations d'évolution dans un espace de Banach quelconque et applications*. Thèse, Orsay, 1972.

[13] H. Brezis, W. A. Strauss. Semi-linear second-order elliptic equations in L^1. *J. Math. Soc. Japan* 25:565-590, 1973.

[14] M. D. Burlică, D. Roşu. A class of nonlinear delay evolution equations with nonlocal initial conditions. *Proc. Amer.* Math. Soc., 142:2445-2458, 2014.

[15] L. Byszewski. Theorem about existence and uniqueness of continuous solution of nonlocal problem for nonlinear hyperbolic equation. *Appl. Anal.* 40:173-180, 1990.

[16] L. Byszewski. Theorems about the existence and uniqueness of solutions of semilinear evolution nonlocal Cauchy problems, *J. Math. Anal. Appl.* 162:494-505, 1991.

[17] L. Byszewski, V. Lakshmikantham. Theorem about the existence and uniqueness of a solution of a nonlocal abstract Cauchy problem in a Banach space. *Appl. Anal.* 40:11-19, 1990.

[18] R. Caşcaval, I. I. Vrabie. Existence of periodic solutions for a class of nonlinear evolution equations. *Rev. Mat. Complut.* 7:325-338, 1994.

[19] O. Cârjă, M. Necula, I. I. Vrabie. *Viability, Invariance and Applications*, Elsevier North-Holland Mathematics Studies 207, Amsterdam, 2007.

[20] D. H. Chen, R. N. Wang, Y. Zhou. Nonlinear evolution inclusions: Topological characterizations of solution sets and applications. *F. Funct. Anal.* 265:2039-2073, 2013.

[21] M. G. Crandall, T. M. Liggett. Generation of semi-groups of nonlinear transformations in general Banach spaces. *Amer. J. Math.* 93:265-298, 1971.

[22] N. Dunford, J. T. Schwartz. *Linear operators. Part I: General theory.* Interscience Publishers Inc. New York, 1958.

[23] R. E. Edwards. *Functional analysis theory and applications*. Holt, Rinehart and Winston, New York Chicago San Francisco Toronto London, 1965.

[24] K. Deng. Exponential decay of solutions of semilinear parabolic equations with initial boundary conditions. *J. Math. Anal. Appl.* 179:630-637, 1993.

[25] J. I. Díaz, I. I. Vrabie. Propriétes de compacité de l'opérateur de Green généralisé pour l'équation des millieux poreux. *C. R. Math. Acad. Sci. Paris* 309:221-223, 1989.

[26] J. I. Díaz, I. I. Vrabie. Existence for reaction diffusion systems: A compactness method approach. *J. Math. Anal. Appl.* 188:521-540, 1994.

[27] Zhenbin Fan, Qixiang Dong, Gang Li. Semilinear differential equations with nonlocal conditions in Banach spaces. *Int. J. Nonlinear Sci.* 2:131-139, 2006.

[28] Ky Fan. Fixed-point and minimax theorems in locally convex topological linear spaces. *Proc. Natl. Acad. Sci. USA* 38:121-126, 1952.

[29] J. García-Falset. Existence results and asymptotic behaviour for nonlocal abstract Cauchy problems. *J. Math. Anal. Appl.* 338:639-652, 2008.

[30] J. García-Falset, S. Reich. Integral solutions to a class of nonlocal evolution equations. *Commun. Contemp. Math.* 12:1032-1054, 2010.

[31] I. L. Glicksberg. A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points. *Proc. Amer. Math. Soc.* 3:170-174, 1952.

[32] D. Gordeziani, M. Avalishvili, G. Avalishvili. On the investigation of one nonclassical problem for Navier-Stokes equations. *AMI* 7:66-77, 2002.

[33] J. Hale. *Functional differential equations*. Applied Mathematical Sciences 3, Springer Verlag, New York, Heidelbreg, Berlin, 1971.

[34] N. Hirano, N. Shioji. Invariant sets for nonlinear evolution equations, Cauchy problems and periodic problems. *Abstr. Appl. Anal.* 2004:183-203, 2004.

[35] M. McKibben. *Discovering Evolution Equations with Applications. Vol. I. Deterministic Models*. Chapman & Hall/CRC Appl. Math. Nonlinear Sci. Ser. London, 2011.

[36] E. Mitidieri, I. I. Vrabie. Existence for nonlinear functional differential equations. *Hiroshima Math. J.* 17:627-649, 1987.

[37] E. Mitidieri, I. I. Vrabie. A class of strongly nonlinear functional differential equations. *Ann. Mat. Pura Appl.* 151:125-147, 1988.

[38] W. E. Olmstead, C. A. Roberts. The One-Dimensional Heat Equation with a Nonlocal Initial Condition. *Appl. Math. Lett.* 10:89-94, 1997.

[39] A. Paicu. Periodic solutions for a class of differential inclusions in general Banach spaces. *J. Math. Anal. Appl.* 337:1238-1248, 2008.

[40] A. Paicu. Periodic solutions for a class of nonlinear evolution equations in Banach spaces. *An. Ştiinţ. Al. I. Cuza Iaşi, (N.S.)* 55:107-118, 2009.

[41] A. Paicu, I. I. Vrabie. A class of nonlinear evolution equations subjected to nonlocal initial conditions. *Nonlinear Anal.* 72:4091-4100, 2010.

[42] F. Rabier, P. Courtier, M. Ehrendorfer. Four-Dimensional Data Assimilation: Comparison of Variational and Sequential Algorithms. *Quart. J. Roy. Meteorol. Sci.* 118:673-713, 1992.

[43] V. V. Shelukhin. A problem nonlocal in time for the equations of the dynamics of a barotropic ocean. *Siberian Math. Journal* 36:701-724, 1995.

[44] I. I. Vrabie. Periodic solutions for nonlinear evolution equations in a Banach space. *Proc. Amer. Math. Soc.* 109:653-661, 1990.

[45] I. I. Vrabie. *Compactness methods for nonlinear evolutions*. Pitman Monographs and Surveys in Pure and Applied Mathematics, Second Edition, London, 1995.

[46] I. I. Vrabie. Existence for nonlinear evolution inclusions with nonlocal retarded initial conditions. *Nonlin. Anal. T.M.A.* 74:7047-7060, 2011.

[47] I. I. Vrabie. Existence in the large for nonlinear delay evolution inclusions with nonlocal initial conditions. *J. Functional Analysis* 262:1363-1391, 2012.

[48] I. I. Vrabie. Nonlinear retarded evolution equations with nonlocal initial conditions. *Dynamic Systems and Applications* 21:417-440, 2012.

[49] I. I. Vrabie. Global solutions for nonlinear delay evolution inclusions with nonlocal initial conditions. *Set-Valued Anal.* 20:477–497, 2012.

[50] I. I. Vrabie. Delay evolution equations with mixed nonlocal plus local initial conditions, *Commun. Contemp. Math.*, **17** (2015) 1350035 (22 pages) DOI: 10.1142/S0219199713500351.

Ann. Acad. Rom. Sci.
Ser. Math. Appl.
ISSN 2066-6594 Vol. 7, No. 1/2015

SEVERAL ITERATIVE PROCEDURES TO COMPUTE THE STABILIZING SOLUTION OF A DISCRETE-TIME RICCATI EQUATION WITH PERIODIC COEFFICIENTS ARISING IN CONNECTION WITH A STOCHASTIC LINEAR QUADRATIC CONTROL PROBLEM*

Vasile Drăgan† Ivan G. Ivanov‡

Dedicated to the memory of Prof. Dr. Viorel Arnăutu

Abstract

We consider a discrete-time periodic generalized Riccati equation. We investigate a few iterative methods for computing the stabilizing solution. The first method is the Kleinman algorithm which is a special case of the classical Newton-Kantorovich procedure, the second one is a method of consistent iterations and two new Stein iterations. The proposed methods are tested and illustrated via some numerical examples.

MSC: 15A24, 15A45, 49N10, 49N20, 65F35

*Accepted for publication on January 5-th, 2015

†vasile.dragan@imar.ro Institute of Mathematics "Simion Stoilow" of the Romanian Academy, Research Unit 2, POBox 1-764, RO-014700, Bucharest, Romania

‡i_ivanov@feb.uni-sofia.bg Faculty of Economics and Business Administration, Sofia University, Sofia 1113, Bulgaria, and Pedagogical College Dobrich, Shoumen University, Shoumen, Bulgaria

Keywords: Discrete-time Riccati equations, periodic coefficients, stabilizing solution, stochastic control, numerical computations.

1 Introduction

Since the pioneer Kalman's work [15], the matrix Riccati differential (difference) equations played a central role in the derivation of the solution of various robust linear quadratic control problems as well as H_2 filtering and H_∞-filtering problems, see e.g. [2, 4, 17] for the continuous-time case, or [3, 9] for the discrete-time case. In [20] where introduced the Riccati differential equations of stochastic control in the case of continuous-time stochastic systems. In the case of discrete-time systems affected by sequences of independent random variables, the discrete-time Riccati equations (DTREs) were introduced in [7, 8, 22].

To solve the linear quadratic optimal control problems on infinite time horizon, a crucial role is played by the so called stabilizing solution of a DTRE. An unified approach of the problem of the existence and uniqueness of a wide class of discrete-time Riccati equations both from deterministic and stochastic framework may be found in the Chapter 5 of [5] for the finite dimensional case and in [19] for infinite dimensional case.

Lately, there is an increasing interest in investigation of several control problems for systems with periodic coefficients. For the readers convenience we refer to [1, 3, 6, 18] and the references therein. Based on the uniqueness of the bounded and stabilizing solution one deduces that in the case of a DTRE with periodic coefficients the bounded and stabilizing solution is also a periodic sequence. This fact is important in the applications because it is necessary finite memory for the offline computation of the gain matrix of the optimal control.

It is worth mentioning that we do not know apriori neither an initial value nor a boundary value of the stabilizing solution of a DTRE. Hence, the existing methods for the computation of a solution with given initial values or boundary values problem for a differential (difference) equation cannot be applied to compute the bounded and stabilizing solution of a DTRE. In the deterministic context there exist two important classes of numerical methods to compute the stabilizing solution of a DTRE namely, the method based on invariant subspaces of associated canonical system [1, 3, 18] and iterative methods [16]. In the case of DTREs from stochastic control the methods based on invariant subspaces are not applicable. Therefore, in this case only iterative methods are mainly used to compute the stabilizing solution of a

Riccati differential (difference) equation. The most popular iterative method is an improved version of Kleinman algorithm. Even if the Kleinman type algorithm is a fast convergent method it has the disadvantage that in the stochastic case require that at each step to compute the unique bounded solution (periodic solution in the periodic case) of a perturbed Lyapunov equation. The numerical computation of such a solution becomes difficult in the case of systems of high dimension of their state space and /or large values of periods in the case of systems with periodic coefficients. That is why in practice were proposed other iterative methods which can be easier implementable (see e.g. Chapter 5 [5] or [14]).

In this paper we consider four iterative methods for computing the stabilizing solution of the discrete-time generalized Riccati equations. There are two Stein iterations which we apply for solving the problem. Similar algorithms for solving the discrete-type algebraic Riccati equations have been developed in our previous investigations [11, 12, 13, 14].

In the last part of the paper, we propose a method to compute the periodic solution occurring at each step of a Kleinman type algorithm. Our method is based on the so called H-representation technique recently developed in [21]. This method allows us to reduce the computation of the periodic solution of a Lyapunov type equation to the computation of the periodic solution of a backward affine equation on an euclidian space of dimension $n(n+1)/2$, n being the dimension of the state space of controlled system under consideration. In the last section of the paper, a comparison between several types of numerical methods discussed in the paper is done.

2 A class of discrete-time Riccati equations of stochastic control (DTRE)

2.1 On the stabilizing solution of DTRE

Consider the discrete-time Riccati equation (DTRE):

$$\begin{aligned} &X(t) = \mathcal{G}(X(t+1)) := \textstyle\sum_{j=0}^{r} A_j^T(t)X(t+1)A_j(t) \\ &-(\textstyle\sum_{j=0}^{r} A_j^T(t)X(t+1)B_j(t) + L(t)) \\ &\times \Big(R(t) + \textstyle\sum_{j=0}^{r} B_j^T(t)X(t+1)B_j(t)\Big)^{-1} \\ &\times(\textstyle\sum_{j=0}^{r} B_j^T(t)X(t+1)A_j(t) + L^T(t)) + M(t), \quad t \in \mathbb{Z}. \end{aligned} \tag{1}$$

This equation arising in connection with the linear quadratic optimization problem described by the discrete-time linear stochastic system:

$$x(t+1) = [A_0(t) + \sum_{j=1}^{r} w_j(t)\, A_j(t)]x(t) + [B_0(t) + \sum_{j=1}^{r} w_j(t)\, B_j(t)]u(t) \quad (2)$$

and the cost functional

$$J(u, x_0) = \sum_{t=0}^{\infty} E\,[\begin{pmatrix} x(t) \\ u(t) \end{pmatrix}^T \begin{pmatrix} M(t) & L(t) \\ L^T(t) & R(t) \end{pmatrix} \begin{pmatrix} x(t) \\ u(t) \end{pmatrix}\,] \qquad (3)$$

with $M(t) = M^T(t), R(t) = R^T(t)$. In (2), $w(t) = (w_1(t), \ldots, w_r(t))^T, t \geq 0$ are independent random vector with zero mean and satisfying $E[w(t)w^T(t)] = I_r$ for all $t \geq 0$. In (2) and (3), $x(t) \in \mathbb{R}^n$ is the state of the system and $u(t) \in \mathbb{R}^m$ are the control parameters.

We make the assumption:

H_1**)** There exists an integer $\theta \geq 1$ such that $A_j(t+\theta) = A_j(t);\ B_j(t+\theta) = B_j(t);\ 0 \leq j \leq r; M(t+\theta) = M(t);\ L(t+\theta) = L(t); R(t+\theta) = R(t),\ t \in \mathbb{Z}$.

Definition 1 *A solution $\{X_s(t)\}_{t\in\mathbb{Z}}$ of DTRE (1) is named stabilizing solution if the zero state equilibrium of the closed-loop system*

$$x(t+1) = [A_0(t) + B_0(t)F_s(t) + \sum_{j=1}^{r} w_j(t)\,(A_j(t) + B_j(t)F_s(t))]x(t) \quad (4)$$

is exponentially stable in mean square (ESMS), where

$$\begin{aligned} F_s(t) \quad &= -\left(R(t) + \textstyle\sum_{j=0}^{r} B_j^T(t)X_s(t+1)B_j(t)\right)^{-1} \\ &\quad \times(\textstyle\sum_{j=0}^{r} B_j^T(t)\, X_s(t)\, A_j(t) + L^T(t)). \end{aligned} \qquad (5)$$

From the developments from Section 5.8 in [5] one deduces a set of necessary and sufficient conditions which guarantee the existence and uniqueness of the bounded and stabilizing solution of DTRE (1).

Proposition 2.1 *Under the assumption H_1**)**, the following are equivalent:*

(i) *DTRE (1) has a unique bounded and stabilizing solution $\{X_s(t)\}_{t\in\mathbb{Z}}$ with the properties:*

(a) *$X_s(\cdot)$ is periodic with period θ;*

(b)

$$R(t)+\sum_{j=0}^{r}B_j^T(t)X_s(t+1)B_j(t)>0,\quad forall\ t\in\mathbb{Z}\,; \tag{6}$$

(ii) *the system (2) is stochastic stabilizable and there exist symmetric matrices* $\hat{X}(t), 0\leq t\leq\theta-1$, *satisfying:*

$$\left(\begin{array}{cc} M(t)-\hat{X}(t) & L(t) \\ L^T(t) & R(t) \end{array}\right)+\sum_{j=0}^{r}(A_j(t)\ \ B_j(t))^T\ \hat{X}(t+1)(A_j(t)\ \ B_j(t))>0 \tag{7}$$

$0\leq t\leq\theta-1$, *with* $\hat{X}(\theta)=\hat{X}(0)$.

Remark 2.1 a) Since any assumption regarding the sign of the quadratic form from (3) was not made, it is not expected to obtain information about the sign of the bounded and stabilizing solution $X_s(\cdot)$. The only relevant information about the solution of the linear quadratic optimization problem described by (2) and (3) is the sign condition (6). In this case, the quadratic part of the discrete-time Riccati equation (1) has defined sign.

b) Even if the stabilizing solution $X_s(\cdot)$ is defined for all $t\in\mathbb{Z}$, from Proposition 2.1 one obtains that under the assumption H_1) it is sufficient to compute a finite number of values $X_s(t), 0\leq t\leq\theta-1$.

The next result may be used to compute a stabilizing control in a state feedback form for the system (2).

Proposition 2.2 *Under the assumption* H_1) *the following are equivalent:*

(i) *the system (2) is stochastically stabilizable;*

(ii) *there exist the matrices* $Y(t)=Y^T(t)>0\in\mathbb{R}^{n\times n}, \Gamma(t)\in\mathbb{R}^{m\times n}, 0\leq t\leq\theta-1$, *satisfying the following system of LMIs:*

$$\left(\begin{array}{cccc} -Y(t) & (\tilde{A}_0(t))^T & \dots & (\tilde{A}_r(t))^T \\ \tilde{A}_0(t) & -Y(t+1) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ \tilde{A}_r(t) & 0 & \dots & -Y(t+1) \end{array}\right)<0 \tag{8}$$

$\tilde{A}_j(t)=A_j(t)Y(t)+B_j(t)\Gamma(t), j=0,\dots,r,\ \ 0\leq t\leq\theta-1$, *with* $Y(\theta)=Y(0)$.

If $(Y(t),\Gamma(t))$, $0\leq t\leq\theta-1$ *is a solution of the LMIs (8), then the control* $u(t)=F(t)\,x(t)$ *stabilizes the system (2), where*

$$F(t)=\Gamma(t-[\frac{t}{\theta}]\theta)\,Y^{-1}(t-[\frac{t}{\theta}]\theta),\quad t\geq 0. \tag{9}$$

(iii) *there exist the matrices* $Y(t) = Y^T(t) \in \mathbb{R}^{n\times n}, \Gamma(t) \in \mathbb{R}^{m\times n}, 0 \le t \le \theta - 1$, *satisfying the following system of LMIs:*

$$\begin{pmatrix} -Y(t+1) & \tilde{A}_0(t) & \dots & \tilde{A}_r(t) \\ (\tilde{A}_0(t))^T & -Y(t) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ (\tilde{A}_r(t))^T & 0 & \dots & -Y(t) \end{pmatrix} < 0 \tag{10}$$

$0 \le t \le \theta - 1$, *with* $Y(\theta) = Y(0)$. *If* $(Y(t), \Gamma(t))$, $0 \le t \le \theta - 1$ *is a solution of the LMIs (10), then the stabilizing feedback gain can be obtained as in (9).*

Proof. One obtains immediately applying Theorem 3.11 and Theorem 3.12 [5] in the case of the corresponding closed-loop systems completed with the Schur complement technique.

2.2 Several iterative procedures to compute the stabilizing solution of DTRE

Here we recall several iterative methods which allow us to compute the bounded and stabilizing solution of DTRE (1).

I. A Newton-Kantorovich type method

For each $k = 1, 2, \dots$ one computes $X^{(k)}(\cdot)$ as the unique periodic solution of the discrete-time backward affine equation:

$$\begin{aligned} X^{(k)}(t) \quad &= \textstyle\sum_{j=0}^{r} (A_j(t) + B_j(t)F^{(k-1)}(t))^T X^{(k)}(t+1) \\ &\quad (A_j(t) + B_j(t)F^{(k-1)}(t)) + Q_{F^{(k-1)}}(t) \end{aligned} \tag{11}$$

where

$$Q_{F^{(k-1)}}(t) = \begin{pmatrix} I_n \\ F^{(k-1)}(t) \end{pmatrix}^T \begin{pmatrix} M(t) & L(t) \\ L^T(t) & R(t) \end{pmatrix} \begin{pmatrix} I_n \\ F^{(k-1)}(t) \end{pmatrix} \tag{12}$$

and

$$\begin{aligned} F^{(k)}(t) \quad &= -\textstyle\sum_{j=0}^{r} \left(R(t) + B_j^T(t) X^{(k)}(t+1) B_j(t) \right)^{-1} \\ &\quad \times (\textstyle\sum_{j=0}^{r} B_j^T(t)\, X^{(k)}(t)\, A_j(t) + L^T(t)) \end{aligned} \tag{13}$$

if $k \ge 1$, while $F^{(0)}(t)$ is a stabilizing feedback gain for the system (2). For example $F^{(0)}(t)$ could be computed via formula (9) either based on a solution of the system of LMIs (8) or a solution of the system of LMIs (10).

One may show in a standard way that if the conditions from Proposition 2.1 (ii) are fulfilled, then for each $k \geq 1$ the control $u(t) = F^{(k)}(t)x(t)$ stabilizes the system (2), thus one obtains that (11) has an unique bounded solution and this solution is periodic with period θ. Furthermore, we have $X^{(k)}(t) \geq X^{(k+1)}(t) \geq \ldots \geq \hat{X}(t),\ k \geq 1, t \in \mathbb{Z},\ \hat{X}(\cdot)$ being any θ-periodic sequence satisfying (7) and $\lim_{k\to\infty} X^{(k)}(t) = X_s(t), t \in \mathbb{Z}$.

Even if the Newton-Kantorovich type method described by (11)-(13) has a quadratic convergence rate it is less used being difficult implementable. The difficulties consist in finding the periodic solution of (11) in the case $r \geq 1$ and $\theta \geq 1$ sufficiently large. That is way ofen alternative methods where derived. Even if those alternative methods have only linear convergence rate, they have the advantage to be easier implementable.

Below, we present some alternative methods to compute the stabilizing solution of DTRE (1). In Section 4 we shall present a method which allows us to compute the θ-periodic solution of (11).

II. A successive approximation method

Step 0. We choose a θ-periodic sequence $\{F^{(0)}(t)\}_{t\in\mathbb{Z}}$ with the property that the control $u(t) = F^{(0)}(t)x(t)$ stabilizes the system (2). For the designing of such a stabilizing feedback gain, may be used, for example, the procedure described by Proposition 2.2. One computes $X^{(1)}(\cdot)$ as a solution of the following system of LMIs:

$$\begin{array}{rl} X^{(1)}(t) & \geq \sum_{j=0}^{r} (A_j(t) + B_j(t)F^{(0)}(t))^T X^{(1)}(t+1) \\ & (A_j(t) + B_j(t)F^{(0)}(t)) + Q_{F^{(0)}}(t) + \varepsilon^2 I_n \end{array} \tag{14}$$

$0 \leq t \leq \theta - 1$, with $X^{(1)}(\theta) = X^{(1)}(0)$, ε is a fixed parameter, $Q_{F^{(0)}}(t)$ being computed as in (12) with $F^{(k-1)}(t)$ replaced by $F^{(0)}(t)$.

Step k, $k \geq 1$. Compute $X^{(k+1)}(\cdot)$ by

$$\begin{array}{rl} X^{(k+1)}(t) & = \sum_{j=0}^{r} (A_j(t) + B_j(t)F^{(k)}(t))^T X^{(k)}(t+1) \\ & (A_j(t) + B_j(t)F^{(k)}(t)) + Q_{F^{(k)}}(t) + \frac{\varepsilon^2}{k+1} I_n\,, \end{array} \tag{15}$$

$Q_{F^{(k)}}(t)$ being computed as in (12) while $F^{(k)}(t)$ is computed as in (13). Since the algorithm described by (14)-(15) is a special case of that described in Section 5.7 from [5], we may conclude that under the conditions of Proposition 2.1 $X^{(1)}(t) \geq \ldots \geq X^{(k)}(t) \geq X^{(k+1)}(t) \geq \ldots \geq \hat{X}(t)$ and $\lim_{k\to\infty} X^{(k)}(t) = X_s(t),\ 0 \leq t \leq \theta - 1$.

In the next section we shall discuss some procedural aspects regarding the computation of a solution of the system of LMIs (14).

III. Stein iterations

Step 0. Coincides with Step 0 from the previous algorithm. One computes $X^{(1)}(\cdot),\ 0 \leq t \leq \theta - 1$ as a solution of the system of LMIs (14). Also one computes $F^{(1)}(t)$ as in (13)) for $k = 1$.

Step k. $k \geq 1$. Compute $X^{(k+1)}(\cdot)$ as a unique $\theta-$periodic solution of the backward Stein equation:

$$\begin{aligned} &X^{(k+1)}(t) \\ &= (A_0(t) + B_0(t)F^{(k)}(t))^T X^{(k+1)}(t+1)(A_0(t) + B_0(t)F^{(k)}(t)) \\ &\quad + \textstyle\sum_{j=1}^{r} (A_j(t) + B_j(t)F^{(k)}(t))^T X^{(k)}(t+1) \\ &\times (A_j(t) + B_j(t)F^{(k)}(t)) + Q_{F^{(k)}}(t) + \frac{\varepsilon^2}{k+1} I_n \end{aligned} \tag{16}$$

$t \in \mathbb{Z}, F^{(k)}(t)$ being computed as in (13).

Some procedural issues regarding the computation of the $\theta-$periodic solution of (16) will be discussed in the next section.

IV. Modified Stein iterations

Step 0. Coincides with Step 0 from the algorithm described in II. One computes $X^{(1)}(\cdot)$ as a solution of the system of LMIs (14) and $F^{(1)}(t)$ as in (13) for $k = 1$.

Step k. $k \geq 1$. One computes $X^{(k+1)}(\cdot)$ as a unique $\theta-$periodic solution of the backward Stein equation:

$$\begin{aligned} &X^{(k+1)}(t) \\ &= (A_0(t) + B_0(t)\Gamma^{(k)}(t))^T X^{(k+1)}(t+1)(A_0(t) + B_0(t)\Gamma^{(k)}(t)) \\ &\quad + \textstyle\sum_{j=1}^{r} (A_j(t) + B_j(t)\Gamma^{(k)}(t))^T X^{(k)}(t+1) \\ &\times (A_j(t) + B_j(t)\Gamma^{(k)}(t)) + Q_{\Gamma^{(k)}}(t) + \frac{\varepsilon^2}{k+1} I_n \end{aligned} \tag{17}$$

where $\Gamma^{(k)}(t) = F_1(t)$ if $k = 1$ and

$$\begin{aligned}\Gamma^{(k)}(t) &= -\big(R(t) + B_0^T(t)\, X^{(k)}(t+1)B_0(t) \\ &\quad + \textstyle\sum_{j=1}^r B_j^T(t)\, X^{(k-1)}(t+1)B_j(t)\big)^{-1} \\ &\quad \times (B_0^T(t)\, X^{(k)}(t+1)A_0(t) \\ &\quad + \textstyle\sum_{j=1}^r B_j^T(t)\, X^{(k-1)}(t+1)A_j(t) + L^T(t))\end{aligned} \tag{18}$$

if $k \geq 2$ and $Q_{\Gamma^{(k)}}(t)$ one computes as in (12) taking $\Gamma^{(k)}(t)$ instead of $F^{(k-1)}(t)$.

3 Procedural issues

In this section we shall analyze some aspects regarding the computation of the sequences of approximations of the stabilizing solution of DTRE (1) described in the previous section.

3.1 The computation of the θ-periodic solution of a backward Stein equation with periodic coefficients

The discrete-time backward affine equations (16)-(17) can be regarded as special cases of the discrete-time backward affine equation:

$$X(t) = \hat{A}^T(t)X(t+1)\hat{A}(t) + H(t) \tag{19}$$

$t \in \mathbb{Z}$, where $\{\hat{A}(t)\}_{t\in\mathbb{Z}} \subset \mathbb{R}^{n\times n}$, $\{H(t)\}_{t\in\mathbb{Z}} \subset \mathcal{S}_n$ are periodic sequences of period θ. Assume that the discrete-time linear equation

$$X(t+1) = \hat{A}(t)X(t) \tag{20}$$

is exponentially stable.
Let $T(t,s) = \hat{A}(t-1)\hat{A}(t-2)...\hat{A}(s)$ if $t > s$ and $T(t,s) = I_n$ if $t = s$, $t, s \in \mathbb{Z}$.

The solutions of equation (19) have the representation:

$$X(t) = T^T(\theta,t)X(\theta)T(\theta,t) + \sum_{s=t}^{\theta-1} T^T(s,t)H(s)T(s,t), t \leq \theta - 1.$$

The periodicity condition $X(0) = X(\theta)$ yields

$$X(\theta) = T^T(\theta,0)X(\theta)T(\theta,0) + \tilde{H} \tag{21}$$

where

$$\tilde{H} = \sum_{s=0}^{\theta-1} T^T(s,0)H(s)T(s,0). \tag{22}$$

Since the zero solution of (20) is exponentially stable it follows that the spectral radius of the monodromy matrix $T(\theta,0)$ satisfies $\rho(T(\theta,0)) < 1$ (see e.g. [3] or [10]).

Hence (21) has a unique solution which may be computed using any existing solver for time invariant Stein equations. Instead of (22), the last term $\tilde{H}$ from (21) may be computed also as: $\tilde{H} = X(0;\theta,0)$ where $t \to X(t;\theta,0)$ is the solution of (19) satisfying the final condition $X(\theta;\theta,0) = 0$. Then, the other values $X(t)$, $1 \le t \le \theta - 1$ of the θperiodic solution of the equation (19) are obtained recursively from this equation.

Remark 3.1. The unique θ-periodic solution of (16) and (17), respectively can be computed according to the procedure described before taking successively $\hat{A}(t) = A_0(t) + B_0(t)F^{(k)}(t)$ in the case of equation (16) or $\hat{A}(t) = A_0(t) + B_0(t)\Gamma^{(k)}(t)$ in the case of equation (17).

3.2 An iterative method for computation of a solution of a system of LMIs (14)

Let $\{F^{(0)}(t)\}_{t\in\mathbb{Z}}$ be a θ-periodic sequence such that the zero solution of the closed-loop system

$$x(t+1) = [A_0(t) + B_0(t)F^{(0)}(t) + \sum_{j=1}^{r} w_j(t)(A_j(t) + B_j(t)F^{(0)}(t))]x(t) \tag{23}$$

is ESMS. Therefore, the discrete-time backward affine equation

$$\begin{aligned} Y(t) \;=\; & \sum_{j=0}^{r}(A_j(t) + B_j(t)F^{(0)}(t))^T Y(t+1)(A_j(t) + B_j(t)F^{(0)}(t)) \\ & + Q_{F^{(0)}}(t) + 2\varepsilon^2 I_n \end{aligned} \tag{24}$$

has a unique θ-periodic solution $\{\tilde{Y}(t)\}_{t\in\mathbb{Z}}$.
Let $Y^{(k)}(t)$ be the θ-periodic solution of the discrete-time backward affine equation:

$$\begin{aligned} Y^{(k)}(t) \;&= [A_0(t) + B_0(t)F^{(0)}(t)]^T Y^{(k)}(t+1) \\ &\quad \times [A_0(t) + B_0(t)F^{(0)}(t)] + H^{(k)}(t) \end{aligned} \tag{25}$$

where

$$H^{(k)}(t) = \sum_{j=1}^{r}(A_j(t)+B_j(t)F^{(0)}(t))^T Y^{(k-1)}(t+1)(A_j(t)+B_j(t)F^{(0)}(t)) + Q_{F^{(0)}}(t) + 2\epsilon^2 I_n, k \geq 1, \quad (26)$$

with

$$Y^{(0)}(t) = 0, t \in \mathbb{Z}. \quad (27)$$

Proposition 3.1. *If the zero solution of (23) is ESMS then the θ-periodic sequences $\{Y^{(k)}(t)\}_{t\in\mathbb{Z}}$, $k = 0, 1, ...$ are well defined via (25)-(27) and have the properties:*

a) $0 = Y^{(0)}(t) \leq Y^{(1)}(t) \leq ... \leq Y^{(k)}(t) \leq ... \leq \tilde{Y}(t)$;

b) $\lim_{k\to\infty} Y^{(k)}(t) = \tilde{Y}(t)$, $t \in \mathbb{Z}$, $\tilde{Y}(\cdot)$ *being the θperiodic solution of 24.*

If k_0 is such that $0 \leq \sum_{j=1}^{r}(A_j(t)+B_j(t)F^{(0)}(t))^T(Y^{(k_0)}(t+1) - Y^{(k_0-1)}(t+1))(A_j(t)+B_j(t)F^{(0)}(t)) \leq \varepsilon^2 I_n$, $0 \leq t \leq \theta-1$, *then* $X^{(1)}(t) \triangleq Y^{(k_0)}(t)$, $0 \leq t \leq \theta-1$, *satisfy the system of LMIs (14).*

The proof is a special case of Corollary 5.3 from [5].

Remark 3.2. For the computation of the θ-periodic solution of the equation (25)-(27) one may use the procedure described in Subsection 3.1.

4 The computation of the θ-periodic solution of a discrete-time backward Stein equation of stochastic control

In this section we shall present an alternative method for the computation of the θ-periodic solution of backward affine equations of type (11)-(13). These equations are special cases of a discrete-time backward affine equation of the form:

$$X(t) = \sum_{j=0}^{r} \hat{A}_j^T(t)X(t+1)\hat{A}_j(t) + G(t) \quad (28)$$

where $\{\hat{A}_j(t)\}_{t\in\mathbb{Z}} \subset \mathbb{R}^{n\times n}$, $0 \leq j \leq r$, $\{G(t)\}_{t\in\mathbb{Z}} \subset \mathcal{S}_n$ are periodic sequences of period θ. Assume that the zero solution of the discrete-time stochastic linear equation:

$$x(t+1) = (\hat{A}_0(t) + \sum_{j=1}^{r} w_j(t)\hat{A}_j(t))x(t) \quad (29)$$

is ESMS. Under these condition (28) has a unique bounded on $\mathbb{Z}$ solution $\hat{X}(\cdot)$ and additionally that solution is periodic with period θ.

Reasoning as in the case of the equation (24) one obtains that $\hat{X}(t) = \lim_{k\to\infty} Z^{(k)}(t)$, where $Z^{(k)}(\cdot), k \geq 1$ is the unique θ-periodic solution of the backward Stein equation:

$$\begin{aligned} Z^{(k)}(t) &= \hat{A}_0^T(t)Z^{(k)}(t+1)\hat{A}_0(t)+\sum_{j=1}^{r}\hat{A}_j^T(t)Z^{(k-1)}(t+1)\hat{A}_j(t)+G(t) \\ Z^{(0)}(t) &= 0, \ \ t \in \mathbb{Z}. \end{aligned} \tag{30}$$

In the following, we shall provide an alternative method which allows us to avoid the iterative process described in (30) to obtain the θ-periodic solution of (28).

4.1 The periodic solution of a discrete-time backward affine equation on an Euclidian space

Let us consider the discrete-time equation

$$x(t) = \hat{M}(t)x(t+1) + g(t) \tag{31}$$

where $\{\hat{M}(t)\}_{t\in\mathbb{Z}} \subset \mathbb{R}^{\hat{n}\times\hat{n}}$, $\{g(t)\}_{t\in\mathbb{Z}} \subset \mathbb{R}^{\hat{n}}$ are periodic sequences of period θ. Assume that the linear equation associated to (31):

$$x(t) = \hat{M}(t)x(t+1) \tag{32}$$

has not nonzero solutions which are periodic of period θ. We set $\hat{T}(t,s) = \hat{M}(t)\hat{M}(t+1)...\hat{M}(s-1)$ if $t < s$ and $\hat{T}(t,s) = I_{\hat{n}}$ if $t = s$. $\hat{T}(t,s)$ is the anti-causal evolution operator defined on $\mathbb{R}^{\hat{n}}$ by the discrete-time backward equation (32).

The solutions of (31) have the representation:

$$x(t) = \hat{T}(t,\tau)x(\tau) + \sum_{s=t}^{\tau-1}\hat{T}(t,s)g(s), \ \forall \ \ t \leq \tau - 1 \in \mathbb{Z}.$$

The periodicity condition $x(0) = x(\theta)$ leads to

$$x(0) = \hat{T}(0,\theta)x(0) + \sum_{s=0}^{\theta-1}\hat{T}(0,s)g(s).$$

Hence, the initial condition $x(0)$ of the unique θ-periodic solution of (31) one obtains solving the system of linear equations

$$(I_{\hat{n}} - \hat{T}(0,\theta))\zeta = \tilde{g} \tag{33}$$

where

$$\tilde{g} = \sum_{s=0}^{\theta-1} \hat{T}(0,s)g(s). \tag{34}$$

Since the linear equation (32) has no nonzero solutions which are periodic sequences of period θ we deduce that $det(I_{\hat{n}} - \hat{T}(0,\theta)) \neq 0$. This allows us to conclude that the equation (33)-(34) has a unique solution $\zeta = \tilde{x}(0) = \tilde{x}(\theta)$. The other values $\tilde{x}(t)$, $1 \leq t \leq \theta - 1$ of the periodic solution $\tilde{x}(\cdot)$ are obtained directly from (31).

Remark 4.1 The term $\tilde{g}$ from (34) may be obtain also from $\tilde{g} = x(0;\theta,0)$ where $ttox(t;\theta,0)$ is the solution of (31) satisfying the final condition $x(\theta;\theta,0) = 0$.

4.2 The H-representation technique revisited

In this paragraph we briefly recall the method of H-representation of a Lyapunov operator in terms of a matrix on the space of dimension $\hat{n} = \frac{n(n+1)}{2}$. This allows us to rewrite the equation (28) in the form of an equation of type (31).

For details we refer to [21], where this method was introduced. We recall that if $X \in \mathbb{R}^{n\times n}$, then $\Psi(X) = Vec(X) = (x(1), x(2), ..., x(n))^T \in \mathbb{R}^{n^2}$ where $x(i)$ is the i^th line of the matrix X, $1 \leq i \leq n$.
Let $E_{11}, E_{12}, ..., E_{1n}, E_{22}, ..., E_{2n}, ..., E_{n-1n-1}, E_{n-1n}, E_{nn}$ be the standard base of the space of symmetric matrices $\mathcal{S}_n$.
This means that $E_{pq} = (e_{pq}(i,j))_{\overline{i,j=1,n}}$ with $e_{pq}(ij) = 1$ if $(ij) \in \{(pq),(qp)\}$ and $e_{pq}(ij) = 0$ otherwise. If $X \in \mathcal{S}_n$ is an arbitrary symmetric matrix, then

$$X = E_{11}x_1 + ... + E_{1n}x_n + E_{22}x_{n+1} + ... + E_{nn}x_{\hat{n}}. \tag{35}$$

We introduce the linear operator $\varphi : \mathcal{S}_n \to \mathbb{R}^{\hat{n}}$ defined by

$$\varphi(X) = x \tag{36}$$

where $x = (x_1, x_2, ..., x_{\hat{n}})^T$ is the vector whose components occur in the right hand side of (35). We introduce also the matrix

$$H = \begin{pmatrix} \Psi(E_{11}) & \Psi(E_{12}) & ... & \Psi(E_{1n}) & \Psi(E_{22}) & ... & \Psi(E_{n-1n}) & \Psi(E_{nn}) \end{pmatrix}.$$

The matrix H has n^2 lines and $\frac{n(n+1)}{2}$ columns. Also, $rankH = \frac{n(n+1)}{2}$. For details see for example [21].
From the definition of the operators φ, Ψ and of the matrix H, we obtain the following fundamental relation:

$$\Psi(X) = H\varphi(X) \tag{37}$$

for all $X \in \mathcal{S}_n$. Let $\mathcal{L}(t) : \mathcal{S}_n \to \mathcal{S}_n$,

$$\mathcal{L}(t)X = \sum_{j=0}^{r} \hat{A}_j^T(t) X A_j(t). \tag{38}$$

Applying Lemma 2.2. in [21] we may write

$$\Psi(\mathcal{L}(t)X) = (\sum_{j=0}^{r} \hat{A}_j^T(t) \otimes \hat{A}_j^T(t))\Psi(X)$$

for all $X \in \mathcal{S}_n$, $\otimes$ being the Kronecker product. Using (37) we obtain

$$\Psi(\mathcal{L}(t)X) = (\sum_{j=0}^{r} \hat{A}_j^T(t) \otimes \hat{A}_j^T(t))H\varphi(X), \ \ \forall X \in \mathcal{S}_n. \tag{39}$$

4.3 The computation of the θ-periodic solution of the equation (28)

Now we show how the computation of the θ-periodic solution of (28) can be reduced to the computation of the θ-periodic solution of an equation of type (31.

First, let us remark that (38) allows us to write (28) in a compact form:

$$X(t) = \mathcal{L}(t)X(t+1) + G(t) \tag{40}$$

Since $\Psi : \mathbb{R}^{n\times n} \to \mathbb{R}^{n^2}$ is an isomorphism we may deduce that the equation (40) is equivalent to the equation:

$$\Psi(X(t)) = \Psi(\mathcal{L}(t)X(t+1)) + \Psi(G(t)). \tag{41}$$

Based on (37) and (39) we rewrite (41) in the form

$$H\varphi(X(t)) = (\sum_{j=0}^{r} \hat{A}_j^T(t) \otimes \hat{A}_j^T(t))H\varphi(X(t+1)) + H\varphi(G(t)). \tag{42}$$

Multiplying to the left (42) by H^T and taking into account that H^TH is invertible, we obtain that $x(t) \triangleq \varphi(X(t))$ is a solution of the discrete-time backward equation on $\mathbb{R}^{\hat{n}}$:

$$x(t) = M(t)x(t+1) + g(t) \tag{43}$$

where

$$M(t) = \sum_{j=0}^{r} (H^TH)^{-1}H^T(\hat{A}_j^T(t) \otimes \hat{A}_j^T(t))H \tag{44}$$

and

$$g(t) = \varphi(G(t)). \tag{45}$$

We have:

Proposition 4.1 *(i) If $\{X(t)\}_{t\in\mathbb{Z}}$ is a global solution of equation (28) then $\{x(t)\}_{t\in\mathbb{Z}}$ defined by $x(t) = \varphi(X(t))$, $t \in \mathbb{Z}$ is a global solution of equation (43)-(45).*

(ii) If $\{\tilde{x}(t)\}_{t\in\mathbb{Z}}$ is a global solution of the backward affine equation (43)-(45), then $\{\tilde{X}(t)\}_{t\in\mathbb{Z}}$ defined by $\tilde{X}(t) = \varphi^{-1}(\tilde{x}(t))$ is a global solution of equation (28).

Proof. (i) follows immediately from the previous developments.

(ii) Let $\tilde{x}(\cdot)$ be a global solution of (43)-(45). If $\tilde{X}(t) = \varphi^{-1}(\tilde{x}(t))$, $t \in \mathbb{Z}$, we define $\Delta(t) = \tilde{X}(t) - \mathcal{L}(t)\tilde{X}(t+1) - G(t)$. We have to show that $\Delta(t) = 0$, for all $t \in \mathbb{Z}$. The previous equality is rewritten as:

$$\tilde{X}(t) = \mathcal{L}(t)\tilde{X}(t+1) + G(t) + \Delta(t). \tag{46}$$

Using again (37), (39) and taking into account that $\varphi(\tilde{X}(t)) = \tilde{x}(t)$ we deduce from (46) that

$$H\tilde{x}(t) = (\sum_{j=0}^{r} \hat{A}_j^T(t) \otimes \hat{A}_j^T(t))H\tilde{x}(t+1) + H\varphi(G(t)) + H\varphi(\Delta(t)). \tag{47}$$

Multiplying to the left (47) by H^T and taking into account that H^TH is invertible, we obtain via (44) and (45) that:

$$\tilde{X}(t) = M(t)\tilde{x}(t+1) + g(t) + \varphi(\Delta(t)).$$

Since $\tilde{x}(\cdot)$ is a solution of (43)-(45) we infer that $\varphi(\Delta(t)) = 0$, $t \in \mathbb{Z}$. Taking into account that φ is an invertible operator, we may conclude that $\Delta(t) = 0$ for all $t \in \mathbb{Z}$, which ends the proof.

Remark 4.2 It is easy to see that $\tilde{x}(t)$, $t \in \mathbb{Z}$, is a θperiodic solution of (43)-(45) if and only if $\varphi^{-1}(\tilde{x}(t))$, $t \in \mathbb{Z}$ is a θ-periodic solution of (28).

Proposition 4.2 *If the zero solution of equation (29) is ESMS then the zero solution is the only one θ-periodic solution of the backward linear equation*

$$x(t) = M(t)x(t+1) \tag{48}$$

associated to (43)-(45).

Proof. Let $\{\hat{x}(t)\}_{t\in\mathbb{Z}}$ be a θ-periodic solution of (48). Let $\hat{X}(t) = \varphi^{-1}(\hat{x}(t))$, $t \in \mathbb{Z}$. From Proposition 4.1 and Remark 4.2 we deduce that $\hat{X}(\cdot)$ is a θ-periodic solution of the linear backward equation

$$X(t) = \mathcal{L}(t)X(t+1). \tag{49}$$

Applying Theorem 2.5 and Theorem 3.11 in [5] we deduce that if the zero-solution of (29) is ESMS, then the discrete-time backward equation (49) has a unique, periodic solution of period θ. Hence, $\hat{X}(t) = 0$, $t \in \mathbb{Z}$. This allows us to deduce that $\hat{x}(t) = \varphi(0) = 0$, $t \in \mathbb{Z}$. Thus the proof is complete.

So, the computation of the value $\tilde{x}(\theta)$ of the θ-periodic solution of the equation (43)-(45) can be performed solving the linear system of $\frac{n(n+1)}{2}$ scalar equations with $\frac{n(n+1)}{2}$ scalar unknowns:

$$(I - T(0,\theta))\zeta = \tilde{g} \tag{50}$$

where

$$\tilde{g} = \sum_{s=0}^{\theta-1} T(0,s)g(s) \tag{51}$$

$T(t,s)$ being the anticausal linear evolution operator on $\mathbb{R}^{\hat{n}}$ defined by the backward linear equation (48) and $g(s)$ are the ones defined in (45).

If $\tilde{x}(\theta) = \zeta$ is the unique solution of the linear equation (50)-(51) then the value $\tilde{X}(\theta)$ of the θ-periodic solution of (28) is obtained by

$$\tilde{X}(\theta) = \varphi^{-1}(\tilde{x}(\theta)). \tag{52}$$

To this end, the components of the vector $\tilde{x}(\theta)$ are plugged in the right hand side of (35). The other related values $\tilde{X}(t)$, $1 \leq t \leq \theta - 1$ of the θ-periodic solution $\tilde{X}(\cdot)$ are obtained directly from (28).

5 Numerical experiments

In this section we present how the considered iterations work for finding a stabilizing solution to (1). We will carry out experiments for numerically solving discrete-time generalized Riccati equation (1).

Our experiments are executed in MATLAB on a 2,16GHz Intel(R) Duo CPU computer. We denote *tol*- a small positive real number denoting the accuracy of computation; $E = max_t\ ||X^{(k)}(t) - \mathcal{G}(X^{(k)}(t+1))||_2$. We use the following stop criterion for all algorithms:

$$E \leq tol\,.$$

5.1 Example 1

Consider a discrete-time 3-periodic linear system with r=1, t=0,1,2, given by (n=3) the coefficient matrices:

$$A_0(0)=\begin{pmatrix} -0.466 & 0.0100 & 0.002 \\ -0.09 & -0.45 & 0.1 \\ -0.035 & -0.01 & -0.485 \end{pmatrix}, \quad A_0(1)=\begin{pmatrix} -0.33 & -0.03 & -0.004 \\ -0.075 & -0.49 & 0.09 \\ -0.025 & -0.015 & -0.495 \end{pmatrix},$$

$$A_0(2)=\begin{pmatrix} -0.45 & 0 & -0.001 \\ -0.095 & -0.505 & 0.1 \\ 0.033 & -0.02 & -0.473 \end{pmatrix}, \quad A_1(0)=\begin{pmatrix} -0.055 & -0.05 & -0.008 \\ 0.13 & -0.12 & 0 \\ -0.3 & 0.25 & 0 \end{pmatrix},$$

$$A_1(1)=\begin{pmatrix} -0.04 & 0.02 & -0.02 \\ 0.2 & -0.035 & -0.01 \\ -0.1 & -0.25 & -0.06 \end{pmatrix}, \quad A_1(2)=\begin{pmatrix} 0 & -0.01 & 0.04 \\ 0.1 & -0.055 & 0 \\ 0.02 & 0.025 & -0.045 \end{pmatrix},$$

$$B_0(0)=\begin{pmatrix} 1 & 12 & -5 \\ 0.1 & -1 & 1.5 \\ 0.2 & -0.5 & 0 \end{pmatrix}, \quad B_0(1)=\begin{pmatrix} 1 & 8 & 4.5 \\ -0.5 & -3 & -2.5 \\ -1 & -0.8 & -0.6 \end{pmatrix},$$

$$B_0(2)=\begin{pmatrix} 1 & -6.5 & -8 \\ 1 & -2.5 & 6 \\ -0.8 & -0.8 & -0.4 \end{pmatrix}, \quad B_1(0)=\begin{pmatrix} -1 & 10 & -5 \\ 0.2 & -1 & -1.5 \\ -0.2 & -2 & -0.5 \end{pmatrix},$$

$$B_1(1)=\begin{pmatrix} 1 & -6.5 & -8 \\ 1 & -2.5 & 6 \\ -0.8 & -0.8 & -0.4 \end{pmatrix}, \quad B_1(2)=\begin{pmatrix} -1 & 10 & -5 \\ 0.2 & -1 & -1.5 \\ -0.2 & -2 & -0.5 \end{pmatrix}.$$

$$L(0)=\tfrac{1}{90}\begin{pmatrix} -0.5 & -0.3 & -0.4 \\ -0.25 & -0.4 & -0.6 \\ -0.5 & -0.5 & -0.8 \end{pmatrix}, \quad L(1)=\tfrac{1}{90}\begin{pmatrix} -0.5 & -0.14 & -0.8 \\ -0.5 & -0.5 & -0.8 \\ -0.6 & -0.8 & -0.3 \end{pmatrix}.$$

$$L(2)=\tfrac{1}{90}\begin{pmatrix} -0.3 & -0.15 & -0.7 \\ -0.6 & -0.6 & -0.5 \\ -0.4 & -0.7 & -0.4 \end{pmatrix}, \quad \begin{cases} R(0) = diag(1.5; 1.5; 1.5), \\ R(1) = diag(1; 1; 1), \\ R(2) = diag(1.25; 1.25; 1.25), \\ M(0) = M(1) = M(2) = 0\,. \end{cases}.$$

We have found the solutions $Y(0), Y(1), Y(2)$ using inequality (8). Then we compute $F(0), F(1), F(2)$ using (9). Thus, we can apply iteration (11). After one iteration steps we obtain the stabilizing solution to (1). The solution is negative definite. Next, we compute the stabilizing solution using iteration (15). We solve inequality (14) for finding $X^{(1)}(0), X^{(1)}(1), X^{(1)}(2)$. We need three LMI iteration steps for solving (14). We find the solution after 7 iteration steps with (15).

Next iteration (16). The solution is obtained after 5 iteration steps.

Next iteration (17). The solution is obtained after 6 iteration steps.

5.2 Two additional examples

Let us consider the new discrete-time 3-periodic linear system with r=1, t=0,1,2. The matrix coefficients are constructed using the following MATLAB code:

$A_j(t) = randn(n, n);\ m1 = max(A_j(t));\ m2 = max(m1);$
$A_j(t) = A_j(t)/(10 * m2); \quad j = 0, 1$
$B_j(t) = randn(n, n);\ \ m1 = max(B_j(t));\ m2 = max(m1);$
$B_j(t) = B_j(t)/(m2); \quad j = 0, 1$
$L(t) = abs(randn(n, n));\ \ m1 = max(L(t));\ m2 = max(m1);$
$L(t) = -L(t)/(80 * m2);$
$M(t) = zeros(n, n);$

5.2.1 Example 2.1

$R(0) = eye(n,n) * 1.05$; $R(1) = eye(n,n) * 0.175$; $R(2) = eye(n,n) * 0.125$.

In this table the full execution time for each iteration is given. This includes the time for computing the initial point $X^{(1)}(0), X^{(1)}(1), X^{(1)}(2)$ and the time for approximating the stabilizing solution using the corresponding iteration formula.

Results for $\mathbf{n = 8}$ and $tol = 1e-5$ for 50 runs are: the CPU time for iteration (11) is 18.0620 seconds; the average number of iteration steps is 2.02 and the maximal error from all runs is $E = 3.7852e-06$.

Results for $\mathbf{n = 8}$ and $tol = 1e-5$ for 50 runs are: the CPU time for iteration (15) is 3.7970 seconds; the average number of iteration steps is 4.1800 and the maximal error from all runs is $E = 4.8106e-06$.

Results for $\mathbf{n = 8}$ and $tol = 1e-5$ for 50 runs are: the CPU time for iteration (16) is 4.6560 seconds; the average number of iteration steps is 4.0 and the maximal error from all runs is $E = 9.0651e-06$.

Results for $\mathbf{n = 8}$ and $tol = 1e-5$ for 50 runs are: the CPU time for iteration (17) is 4.77 seconds; the average number of iteration steps is 5.06 and the maximal error from all runs is $E = 6.9399e-06$.

Results for $\mathbf{n = 12}$ and $tol = 1e-5$ for 50 runs are: the CPU time for iteration (11) is 125.9060 seconds; the average number of iteration steps is 2.08 and the maximal error from all runs is $E = 9.5401e-06$.

Results for $\mathbf{n = 12}$ and $tol = 1e-5$ for 50 runs are: the CPU time for iteration (15) is 11.7350 seconds; the average number of iteration steps is 4.26 and the maximal error from all runs is $E = 9.4270e-06$.

Results for $\mathbf{n = 12}$ and $tol = 1e-5$ for 50 runs are: the CPU time for iteration (16) is 12.8440 seconds; the average number of iteration steps is 4.160 and the maximal error from all runs is $E = 8.4240e-06$.

Results for $\mathbf{n = 12}$ and $tol = 1e-5$ for 50 runs are: the CPU time for iteration (17) is 15.0380 seconds; the average number of iteration steps is 5.36 and the maximal error from all runs is $E = 9.0533e-06$.

5.2.2 Example 2.2

We choose: $R(0) = eye(n,n) * 1.45$; $R(1) = eye(n,n) * 0.175$; $R(2) = eye(n,n) * 0.125$.

We present the full information about each iteration. This includes the time for approximating the stabilizing solution using the corresponding iteration formula. The initial point $X^{(1)}(0), X^{(1)}(1), X^{(1)}(2)$ is the same for

all iterations and it is computing via (14).

Results for $\mathbf{n = 18}$ and $tol = 1e-4$ for 50 runs are: the CPU time for iteration (11) is 811.5010 seconds; the average number of iteration steps is 2.36 and the maximal error from all runs is $E = 9.8986e-05$.

Results for $\mathbf{n = 18}$ and $tol = 1e-4$ for 50 runs are: the CPU time for iteration (15) is 0.622 seconds; the average number of iteration steps is 4.5800 and the maximal error from all runs is $E = 8.2157e-05$.

Results for $\mathbf{n = 18}$ and $tol = 1e-4$ for 50 runs are: the CPU time for iteration (16) is 3.312 seconds; the average number of iteration steps is 4.2200 and the maximal error from all runs is $E = 9.3210e-05$.

Results for $\mathbf{n = 18}$ and $tol = 1e-4$ for 50 runs are: the CPU time for iteration (17) is 3.61 seconds; the average number of iteration steps is 5.46 and the maximal error from all runs is $E = 8.6559e-05$.

6 Conclusion

We have considered four iterations for computing the stabilizing solution to (1). In order to execute iterations (11), (16) and (17) we have to solve a linear system with a big dimension, i.e. it has the size $(\theta\, n)^2 \times (\theta\, n)^2$. In the same time iteration (15) gives us a possibility to find $X^{(k+1)}(t)$ in the following way:

$$
\begin{aligned}
&X^{(k+1)}(\theta-2) = \\
&= \textstyle\sum_{j=0}^{r} (A_j(\theta-2) + B_j(\theta-2)F^{(k)}(\theta-2))^T X^{(k)}(\theta-1) \\
&\times (A_j(\theta-2) + B_j(\theta-2)F^{(k)}(\theta-2)) + Q_{F^{(k)}}(\theta-2) + \tfrac{\varepsilon^2}{k+1}\, I_n
\end{aligned}
$$

$$
\begin{aligned}
&X^{(k+1)}(\theta-3) = \\
&= \textstyle\sum_{j=0}^{r} (A_j(\theta-3) + B_j(\theta-3)F^{(k)}(\theta-3))^T X^{(k+1)}(\theta-2) \\
&\times (A_j(\theta-3) + B_j(\theta-3)F^{(k)}(\theta-3)) + Q_{F^{(k)}}(\theta-3) + \tfrac{\varepsilon^2}{k+1}\, I_n
\end{aligned}
$$

....

$$
\begin{aligned}
&X^{(k+1)}(0) = \\
&= \textstyle\sum_{j=0}^{r} (A_j(0) + B_j(0)F^{(k)}(0))^T X^{(k+1)}(1)(A_j(0) + B_j(0)F^{(k)}(0)) \\
&\quad + Q_{F^{(k)}}(0) + \tfrac{\varepsilon^2}{k+1}\, I_n
\end{aligned}
$$

$$X^{(k+1)}(\theta-1) =$$
$$= \sum_{j=0}^{r} \left(A_j(\theta-1) + B_j(\theta-1)F^{(k)}(\theta-1)\right)^T X^{(k+1)}(0)$$
$$\times(A_j(\theta-1) + B_j(\theta-1)F^{(k)}(\theta-1)) + Q_{F^{(k)}}(\theta-1) + \frac{\varepsilon^2}{k+1} I_n$$

We call the last iteration the improved approximation method. This method is applied to Example 2.2. Results for $\mathbf{n = 18}$ and $tol = 1e-4$ for 50 runs are: the CPU time is 0.02 seconds; the average number of iteration steps is 2.06 and the maximal error from all runs is $E = 3.7096e-05$.

Acknowledgement. The work of the first author was supported by Grant 145/2011 of CNCS Romania.

References

[1] F.A. Aliev, V.B. Larin, Optimization Problems for Periodic Systems, *Int. Appl. Mech.*,45,11, 1162?188, 2009.

[2] B.D.O. Anderson, J.B. Moore, *Optimal Control: Linear Quadratic Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[3] S. Bittanti, P.Colaneri. *Periodic Systems, Filtering and Control*, Springer- Verlag, London, 2009.

[4] V. Drăgan, A. Halanay, *Stabilization of linear systems* - Birkhauser, Boston, 1999.

[5] V. Drăgan, T. Morozan, A.M.Stoica. *Mathematical Methods in Robust Control of Discrete-time Linear Stochastic Systems.* Springer, New-York, 2010.

[6] V. Drăgan, S. Aberkane, I. G. Ivanov, On computing the stabilizing solution of a class of discrete-time periodic Riccati equations, *Int. J. Robust Nonlinear Control*, (2013), published online DOI: 10.1002/rnc.3131.

[7] A. Halanay, T. Morozan, Stabilization by linear feedback of linear discrete stochastic systems, *Rev. Roumaine Math. Pures Appl.,*23,4, 561-571, 1978.

[8] A. Halanay, T. Morozan, Optimal stabilizing compensators for linear discrete-time systems under independent perturbations,*Revue Roum. Math. Pure et Appl.*, 37, 3, 213-224, 1992.

[9] A. Halanay, V. Ionescu, *Time varying discrete linear systems,* Berlin, Birkhauser, 1994.

[10] A. Halanay, Vl. Rasvan. *Discrete Time Systems. Stability and Stable Oscillations*, Gordon and Breach, 2000.

[11] I. Ivanov, Accelerated LMI solvers for the maximal solution to a set of discrete-time algebraic Riccati equations, *Appl. Math. E-Notes.* 12: 228–238, 2012. http://www.math.nthu.edu.tw/ amen/, open access

[12] I. Ivanov, Iterations for a General Class of Discrete-Time Riccati-Type Equations: A Survey and Comparison, open access, DOI: 10.5772/45718, 2012.

[13] I.Ivanov, An Improved Method for Solving a System of Discrete-Time Generalized Riccati Equations, *Journal of Numerical Mathematics and Stochastics.* 3(1): 57-70, 2011. http://www.jnmas.org/jnmas3-7.pdf

[14] I.Ivanov, V. Dragan, Decoupled Stein Iterations to the Discrete-time Generalized Riccati Equations, *IET Control Theory Appl.* 6(10): 1400–1409, 2012.

[15] R.E. Kalman, Contributions to the theory of optimal control, *Bull. Soc. Math. Mex.*, 5, 102?19, 1960.

[16] D.L. Kleinman, On an iterative technique for Riccati equation computation, *IEEE Transactions on Automatic Control*, 13, 114?15, 1968.

[17] H. Kwakernaak, R. Sivan, *Linear optimal control systems*, New York, Willey Interscience, 1972.

[18] V.B. Larin, High-Accurasy Algorithms for Solving of Discrete Periodic Riccati Equation, *Appl. Comput. Math.*, 6, 1, 10-17, 2007.

[19] V. M. Ungureanu, V. Drăgan, T. Morozan, Global solutions of a class of discrete-time backward nonlinear equations on ordered Banach spaces with applications to Riccati equations of stochastic control, *Optimal Control, Applications and Methods*, 34, 2, 164–190, 2013.

[20] W.M. Wonham, Random differential equations in control theory, *in Probabilistic Methods in Applied Mathematics, vol. 2 (Academic, New York)*, pp. 131?12, 1970.

[21] W. Zhang, B. Chen. H-Representation and Applications to Generalized Lyapunov Equations and Linear Stochastic Systems. *IEEE Trans. on Aut. Contr.*57 (12): 3009–3022, 2012.

[22] J. Zabczyk, Stochastic control of discrete-time systems, *Control Theory and Topics in Funct. Analysis*, 3, IAEA, Vienna, 1976.

Ann. Acad. Rom. Sci.
Ser. Math. Appl.
ISSN 2066-6594 Vol. 7, No. 1/2015

UNILATERAL CONDITIONS ON THE BOUNDARY FOR SOME SECOND ORDER DIFFERENTIAL EQUATIONS*

Dan TIBA†

Dedicated to the memory of Prof. Dr. Viorel Arnăutu

Abstract

Sufficient conditions for the existence of solutions in strongly nonlinear boundary value problems of elliptic and parabolic type, including ordinary differential equations with unilateral conditions on the boundary, are derived by means of an abstract scheme for continuous perturbations of accretive operators in Banach spaces.

MSC: 35J25, 35K20, 34B15

keywords: nonlinear boundary value problems, partial differential equations, ordinary differential equations

1 Introduction

This paper is concerned with strongly nonlinear boundary value problems of elliptic type

$$Au + f(x, u, \text{grad } u) = 0, \quad a.e.\ \Omega \tag{1}$$

*Accepted for publication on January 15-th, 2015

†Institute of Mathematics "Simion Stoilow" (Romanian Academy) and Academy of Romanian Scientists, Bucharest; dan.tiba@imar.ro

$$-\frac{\partial u}{\partial \nu} \in \beta(u) \quad a.e.\ \Gamma \tag{2}$$

and parabolic type

$$\frac{\partial u}{\partial t} = Au + f(x, u, \text{grad } u) = 0 \quad a.e.\ ]0,T[\times\Omega \tag{3}$$

$$u(0) = u_0 \quad a.e\ \Omega \tag{4}$$

$$-\frac{\partial u}{\partial \nu} \in \beta(u) \quad a.e.\ [0,T] \times \Gamma.$$

We also prove some existence results for the two points and the periodic problem associated with ordinary differential equations

$$-u''(s) + f(s, u(s), u'(s)) = 0 \quad a.e.\ [0,1], \tag{5}$$

which can be compared with the classical result of Bernstein [7].

Above A denotes a second order elliptic operator, f is function satisfying the Caratheodory assumptions, β is the subdifferential of a convex, lower-semicontinuous, proper function $j : R \to]-\infty, +\infty]$ and Ω is a bounded domain in R^N with sufficiently smooth boundary Γ.

The following notation will be used throughout this paper. If E is a Banach space, we shall denote by $L^p(0,T;E),\ 1 \leq p \leq \infty$, the space of all p - integrable, E - valued functions on $[0,T]$ and by $C(0,T;E)$ the Banach space of all continuous functions from $[0,T]$ to E. We shall denote by $W^{1,p}(0,T;E)$ the space of all p - integrable, E - valued distributions y with derivative y' taken in the sense of vectorial distributions on $]0,T[$, p - integrable. Equivalently, $y' \in W^{1,p}(0,T;E)$ means that $y : [0,T] \to E$ is absolutely continuous, almost everywhere differentiable on $]0,T[$ and $y' \in L^p(0,T;E)$. By $W^{k,p}(\Omega)$ we mean the usual Sobolev space of real distributions in Ω. We shall use the symbols $||\cdot||_p$, $||\cdot||_{k,p}$ for the norms in $L^p(\Omega)$, $W^{k,p}(\Omega)$ respectively. In the case $p = 2$, we put $H^k(\Omega)$ instead of $W^{k,p}(\Omega)$.

We assume familiarity with concepts and methods of nonlinear monotone equations and we refer to Barbu [2], Brezis [3], [4] for significant results in this field. However, for easy references we recall some facts about subdifferentials.

Let $\varphi : E \to]-\infty, +\infty]$ be a convex, lower semicontinuous, proper function. We denote by $\partial\varphi(x)$ the set of all $z \in E'$, the dual of E, such that

$$\varphi(x) \leq \varphi(y) + (x - y, z) \quad \forall\, y \in E,$$

and call it the subdifferential of φ at x, where $(\cdot,\cdot)$ is the pairing between E and E'.

Conditions of type (2), (4) are called unilateral conditions on the boundary and they arise in elasticity. See for instance Duvaut - Lions [10], Goeleven [11], Goeleven et. al. [12].

Problems (1) - (4) are very much discussed in the literature. We mention the papers for Brezis - Haraux [5], Brezis - Nirenberg [6], Vy Khoi Le [14] that deal with the case when the nonlinear term f does not depend on grad u or the elliptic operator is degenerate and with Landesman - Lazer conditions.

Equations of form (1), (3) appear in the paper of Puel [16], but the problem is the Dirichlet one with unilateral constraints in the interior of Ω and certainly the methods are different. Our method of proof is similar to that used in [19], [20]. Regularity results and various extensions are discussed in [8], [9], [13], [18].

Our approach applies to a large class of problems and, in certain cases, quadratic growth with respect to the gradient is allowed.

In the subsequent sections we introduce an abstract scheme based on m - accretive operators and we apply it to elliptic, parabolic and ordinary differential boundary value problems. An Appendix briefly analyzes some properties of the Nemitsky operator.

2 An Abstract Perturbation Scheme

Let W be a Banach space, topologically and algebraically included in X, another Banach space with dual X' uniformly convex.

Proposition 1. *Let $T : X \to X$ be a m - accretive operator with $0 \in T0$, $D(T) \subset W$ and $(\lambda I + T)^{-1} : X \to W$ compact for some $\lambda \geq 0$. Let $S : W \to X$ be a bounded, demicontinuous mapping.*

Then, for every $m \in N$, there is $x_m \in W$, such that

$$\lambda x_m + Tx_m + S_m x_m \ni 0. \tag{6}$$

Here we have defined the truncate S_m of $S - \lambda I$ by

$$S_m x = \begin{cases} Sx - \lambda x, & ||x||_W \leq m \\ S(\dfrac{mx}{||x||_W}) - \lambda \dfrac{mx}{||x||_W}, & ||x||_W > m \end{cases}.$$

Proof.

The equation (6) can be written as

$$x_m = (\lambda I + T)^{-1}(-S_m x_m).$$

The operator defined by the right hand side is compact in W because $(\lambda I + T)^{-1}$ is and S_m is bounded. It maps a certain sphere with a sufficiently large radius in itself because S_m is uniformly bounded on W and $(\lambda I + T)^{-1}$ is compact from X in W.

It is continuous. Here is the argument

Let $x_n \to x$ in W, then $S_m x_n \to S_m x$ weakly in X because S_m is also demicontinuous. It yields $\{S_m x_n\}_n$ to be bounded in X that is extracting a convenient subsequence, denoted again by x_n, we have $(\lambda I + T)^{-1} \cdot (-S_m x_n) \to y$ strongly in W. Hence $(\lambda I + T)^{-1}(-S_m x_n) \to y$ strongly in X. Operator $(\lambda I + T)^{-1}$ is single-valued, demiclosed in X, so $(\lambda I + T)^{-1} \cdot (-S_m x) = y$.

Therefore, one can use the Schauder fixed point theorem to obtain the desired solution.

3 Elliptic Problems

Let A be the second order elliptic operator

$$Au = -\sum_{i,j} \frac{\partial}{\partial x_j}(a_{ij}\frac{\partial u}{\partial x_i}) + u \tag{7}$$

with

$$\sum_{i,j} a_{ij}(x)\xi_i\xi_j \geq \alpha|\xi|^2 \quad \text{a.e. } \Omega,\ \alpha > 0,\ \xi \in R^N. \tag{8}$$

Here $a_{ij} \in C^1(\Omega)$, $a_{ij} = a_{ji}$ and Ω is a bounded domain with a sufficiently smooth boundary Γ.

We denote by $\dfrac{\partial u}{\partial \nu}$ the conormal derivative associated to A

$$\frac{\partial u}{\partial \nu} = \sum_{i,j} a_{ij}\frac{\partial u}{\partial x_i}\cos(\bar{n}, x_j) \tag{9}$$

where $\bar{n}$ is the exterior normal to Ω.

Consider two real numbers $2 \geq q \geq 1,\ p > 1$ such that $W^{1,q}(\Omega) \subset L^p(\Omega)$ topologically and $W^{2,p}(\Omega) \subset W^{1,q}(\Omega)$ with compact inclusion. The existence of these numbers is ensured by the wellknown Sobolev embedding theorem.

Assume that $f : \Omega \times R \times R^N \to R$ satisfies the Caratheodory conditions

$f(\cdot, u, v_1, \ldots, v_N)$ is measurable for every u, v.

$f(x, \cdot, \cdot)$ is continuous a.e. $x \in \Omega$.

The Nemitsky operator $S : W^{1,q}(\Omega) \to L^p(\Omega)$ defined by

$$(Su)(x) = f(x, u(x), \text{grad } u(x)) \quad \text{a.e. } \Omega \tag{10}$$

satisfies

$$S \textit{ is bounded} \tag{11}$$

$$S \textit{ is demicontinuous.} \tag{12}$$

See the Appendix for a discussion of such hypotheses. Moreover, the following growth restriction is needed

$$f(x, u, v)u \geq K|u|^s - d|v|^2 - \gamma(x) \cdot u \tag{13}$$

where $K > 0$, $s > 1$ is choosen such that $W^{1,q}(\Omega) \subset L^s(\Omega)$, $\gamma \in L^\infty(\Omega)$ and d is a small constant.

Remark 1. Condition $W^{2,p}(\Omega) \subset W^{1,q}(\Omega)$, with compact inclusion, shows that growth order of $f(x, u, \cdot)$, which is $\frac{q}{p}$ (see Appendix), cannot exceed 2 when $N = 2$, cannot exceed $\frac{3}{2}$ when $N = 3$ and so on, according to the Sobolev inequalities.

Remark 2. We give a simple example of function $f(x, u, v)$, where $v = (v_1, \ldots, v_N) \in R^N$

$$f(x, u, v) = |u|^r \cdot u + |v|^{\frac{q}{p}} \eta(u) + \gamma(x)$$

Conditions (11), (12), (13) are fulfilled evidently for an appropriate r (see the Appendix) under assumption that $\eta : R \to R$ is a monotone continuous and bounded function.

Remark 3. The operator A can be more generally

$$A'u = -\sum_{i,j} \frac{\partial}{\partial x_j}(a_{ij} \frac{\partial u}{\partial x_i}) + \sum_i b_i \frac{\partial u}{\partial x_i} + cu.$$

The last two terms can be taken in $f(x, u, v)$ and one can apply the present results under appropriate conditions on b_i, $c > 0$, [4], p. 6.

Theorem 1. *Under the above hypotheses, problem (1), (2) has at least one solution u in $W^{2,p}(\Omega)$.*

Proof.

We apply *Proposition 1* with $W^{1,q}(\Omega)$ and $X = L^p(\Omega)$.

Operator $T : L^p(\Omega) \to L^p(\Omega)$ defined by

$Tu = Au$

$D(T) = \{u \in W^{2,p}(\Omega); Au \in L^p(\Omega), -\dfrac{\partial u}{\partial \nu} \in \beta(u)\}$

is m - accretive and $(T+\lambda I)^{-1}$ is bounded from $L^p(\Omega)$ in $W^{2,p}(\Omega)$ for $\lambda > 0$ large enough, according to Brezis [4], Proposition I.13 and Remark I.22. It yields that $(T + \lambda I)^{-1}$ is compact operator from $L^p(\Omega)$ in $W^{1,q}(\Omega)$.

Then for every natural number m, there is u_m in $W^{2,p}(\Omega)$ such that

$$\lambda u_m + Tu_m + S_m u_m \ni 0. \tag{14}$$

We assume that $||u_m||_{1,q} > m$, otherwise u_m satisfies (1), (2) and the problem is solved.

Equation (14) becomes

$$\lambda u_m + Au_m + f(x, \frac{mu_m}{||u_m||_{1,q}}, \frac{m \text{ grad } u_m}{||u_m||_{1,q}}) - \lambda \frac{mu_m}{||u_m||_{1,q}} \ni 0. \tag{15}$$

Multiply by u_m and integrate over Ω

$$\int_\Omega Au_m \cdot u_m dx + \int_\Omega f(x, \frac{mu_m}{||u_m||_{1,q}}, \frac{m \text{ grad } u_m}{||u_m||_{1,q}}) u_m dx \leq 0.$$

Integrating by parts, using (8) and (2) we get

$$\alpha ||u_m||_{1,2} + \int_\Omega f(x, \frac{mu_m}{||u_m||_{1,q}}, \frac{m \text{ grad } u_m}{||u_m||_{1,q}}) u_m dx \leq 0.$$

From (13) one obtains $\{u_m\}$ to be bounded in $H^1(\Omega)$, that is for m large enough u_m verifies (1), (2) and the proof is finished.

Remark 4. Not only classical problems, but many boundary problems can be expressed in form (2).

Example 1. Let $j : R \to]-\infty, +\infty]$ be a convex, lower semicontinuous, proper function given by

$$j(s) = \begin{cases} 0 & \text{if } s = 0 \\ +\infty & \text{otherwise} \end{cases}.$$

Then $\beta = \partial j$ is

$$\beta(s) = \begin{cases} \mathbb{R} & \text{if } s = 0 \\ \emptyset & \text{otherwise} \end{cases} .$$

and condition (2) is the Dirichlet one.

Example 2. Let $j(s) = 0$ for every s. Then $\beta(s) = 0$ for every s and condition (2) corresponds to the Neumann problem.

Example 3. Consider

$$j(s) = \begin{cases} 0 & s \geq 0 \\ +\infty & s < 0 \end{cases} .$$

Then

$$\beta(s) = \begin{cases} 0 & s > 0 \\ ]-\infty, 0] & s = 0 \\ \emptyset & s < 0 \end{cases} .$$

We obtain for (2) the Signorini boundary conditions.

Example 4. Consider $j(s) = |s|$. In this case

$$\beta(s) = \operatorname{sgn}(s) = \begin{cases} 1 & s > 0 \\ [-1, 1] & s = 0 \\ -1 & s < 0 \end{cases} .$$

The corresponding condition (2) appears in elasticity.

4 Parabolic Problems

For the sake of simplicity we take the problem

$$\frac{\partial u}{\partial t} - \Delta u + f(x, u, \operatorname{grad} u) = 0 \quad \text{a.e. }]0, T[\times\Omega \tag{16}$$

$$u(0, x) = u_0(x) \quad \text{a.e. } \Omega \tag{17}$$

$$-\frac{\partial u(t, x)}{\partial n} \in \beta(u(t, x)) \quad \text{a.e. } [0, T] \times \Gamma. \tag{18}$$

We start with the following lemma

Lemma 1 *The operator* $B : L^2(0,T;L^2(\Omega)) \to L^2(0,T;L^2(\Omega))$ *defined by*

$$Bu = \frac{\partial u}{\partial t} - \Delta u$$

$$D(B) = \{u \in H^1(0,T;H^2(\Omega)); u(0,x) = u_0(x), -\frac{\partial u(t,x)}{\partial n} \in \beta(u(t,x))\}$$

is maximal monotone and for $u_0 \in D(\varphi)$, B^{-1} *is compact from* $L^2(0,T;L^2(\Omega))$ *in* $L^2(0,T;H^1(\Omega))$.

Here $\varphi : L^2(\Omega) \to]-\infty, +\infty]$ is a proper, lower-semicontinuous, convex function given by

$$\varphi(u) = \begin{cases} \frac{1}{2}\int\limits_\Omega |\text{grad } u|^2 dx + \int\limits_\Gamma j(u) d\tau & \text{if } u \in H^1(\Omega),\ j(u) \in L^1(\Gamma) \\ +\infty & \text{otherwise} \end{cases}$$

and $\partial\varphi = -\Delta$ with

$$D(\partial\varphi) = \{u \in H^2(\Omega); -\frac{\partial u}{\partial n} \in \beta(u) \text{ a.e. } \Gamma\}.$$

Proof

One easily can check, using the Green formula, that operator B is monotone. To obtain the maximality it suffices that problem

$$\frac{\partial u}{\partial t} - \Delta u + u(t,x) = \tilde{f}(t,x) \quad \text{a.e. } \Omega\times]0,T[\tag{19}$$

$$u(0,x) = u_0(x) \quad \text{a.e. } \Omega \tag{20}$$

$$-\frac{\partial u(t,x)}{\partial n} \in \beta(u(t,x)) \quad \text{a.e. }]0,T[\times\Gamma \tag{21}$$

has at least one solution for every $\tilde{f} \in L^2(0,T;L^2(\Omega))$.

Operator $Cu = -\Delta u + u$ with

$$D(C) = \{u \in H^2(\Omega); -\frac{\partial u}{\partial n} \in \beta(u)\}$$

is a subdifferential (see Barbu [2], p. 63).

Therefore, we can apply the smoothing effect on initial data and problem (10) - (12) has at least one solution u for every $f \in L^2(0,T;L^2(\Omega))$ and $u_0 \in L^2(\Omega)$ (see Barbu [2], p. 189).

If $u_0 \in D(\varphi)$ then $\dfrac{\partial u}{\partial t} \in L^2(0,T;L^2(\Omega))$, $\Delta u \in L^2(0,T;L^2(\Omega))$ and the mapping $\tilde{f} \to u$ is compact from $L^2(0,T;L^2(\Omega))$ in $L^2(0,T;W^{1,2}(\Omega))$ in the case $u_0 \in D(\varphi)$ and the proof is finished.

Assume now that $f : \Omega \times R \times R^N \to R$ satisfies the Caratheodory conditions and operator $S : L^2(0,T;H^1(\Omega)) \to L^2(0,T;L^2(\Omega))$ defined by

$$(Su)(t,x) = f(x, u(t,x),\ \mathrm{grad}_x u(t,x))$$

satisfies hypotheses (11) - (13) with $p = q = 2$.

One can state

Theorem 2 *Under the above hypotheses, problem (16) - (18) has at least one solution u in $H^1(0,T;H^2(\Omega))$.*

Proof

According to *Lemma 1*, we can apply *Proposition 1* with $\lambda = 0$, $X = L^2(0,T;L^2(\Omega))$, $W = L^2(0,T;H^1(\Omega)))$ and obtain the approximate equations

$$\frac{\partial u_m}{\partial t} - \Delta u_m + S_m u_m \ni 0.$$

Suppose that the norm of u_m in $L^2(0,T;H^1(\Omega))$ denoted $||u_m||_W$ strictly exceeds m, for every natural number m.

The approximate equations become

$$\frac{\partial u_m}{\partial t} - \Delta u_m + f(x, \frac{m u_m}{|||u_m||_W}, \frac{m\ \mathrm{grad}_x u_m}{||u_m||_W}) \ni 0. \tag{22}$$

Multiply by $u_m(s,x)$ and integrate over $[0,t]$

$$\frac{1}{2}|u_m(t,x)|^2 - \frac{1}{2}|u_0(x)|^2 - \int_0^t \Delta u_m(s,x) \cdot u_m(s,x) ds +$$

$$+ \int_0^t f(x, \frac{m u_m}{|||u_m||_W}, \frac{m\ \mathrm{grad}_x u_m}{||u_m||_W}) \cdot u_m ds = 0.$$

Integrating over Ω, using the Green formula, we get

$$\frac{1}{2}\int_\Omega |u_m(t,x)|^2 dx + \int_0^t \int_\Omega |\mathrm{grad}_x u_m(s,x)|^2 dx ds) + \tag{23}$$

$$+\int_0^t\int_\Omega f(x,\frac{mu_m}{|||u_m||_W},\frac{m\ \mathrm{grad}_x u_m}{||u_m||_W})\cdot u_m dxds\le C.$$

From condition (13), when $t=T$ it yields u_m to be bounded in $L^2(0,T;H^1(\Omega))$ and using again (23) we see that u_m is bounded in $C(0,T;H^1(\Omega))$. Then for large m we have $||u_m||_W\le m$, that is u_m satisfies problem (16) - (18). The regularity is obtained as in (19) - (21).

5 Ordinary Differential Equations

We take into account the two point boundary value problem:

$$-u''(t)+f(t,u(t),u'(t))=0\quad \text{a.e. } t\in[0,1] \tag{24}$$

$$u(0)=a,\ u(1)=b \tag{25}$$

where f is Caratheodory:

- $f(t,u,v)$ measurable in t for every u,v
- $f(t,u,v)$ continuous in u,v a.e. $t\in[0,1]$

and a,b are real numbers.

We assume that

$$|f(t,u,v)|\le g(t,u)+h(t,u)|v|^2 \tag{26}$$

with

$$\sup_{|u|\le r}|g(t,u)|\in L^2(0,1)$$

$$\sup_{|u|\le r}|h(t,u)|\in L^\infty(0,1)$$

for every $r>0$, and

$$f(t,u,v)\cdot u\ge K(u)\cdot v-\alpha|u|^2+\gamma,\ \alpha<1 \tag{27}$$

where K is a continuous, d - homogeneous function, that is

$$K(\lambda u)=\lambda^d K(u),\ \lambda>0,\ d\ge 0.$$

Theorem 3 *Under the above hypotheses, problem (24), (25) has at least one solution u in $W^{2,1}(0,1)$.*

Proof

Shifting the domain of f in u, v one can suppose instead of (25)

$$u(0) = u(1) = 0 \tag{28}$$

(null Dirichlet boundary conditions).

Operator $T : L^2(0,1) \to L^2(0,1)$ defined by

$Tu = -u''$

$D(T) = \{u \in H^2(0,1); u(0) = u(1) = 0\}$

is maximal monotone and $(I+T)^{-1}$ is compact from $L^2(0,1)$ in $W^{1,4}(0,1)$.

Under condition (26) operator $S : W^{1,4}(0,1) \to L^2(0,1)$ defined by $(Su)(t) = f(t, u(t), u'(t))$ is bounded and continuous (see the Appendix).

We can use *Proposition 1* with $\lambda = 1$, $X = L^2(0,1)$, $W = W^{1,4}(0,1)$ to derive the existence of approximating solutions

$$u_m(t) - u_m''(t) + S_m u_m(t) = 0.$$

Assume that $||u_m||_{1,4} > m$ for every m. Then

$$u_m(t) - u_m''(t) + f(t, \frac{mu_m(t)}{|||u_m||_{1,4}}, \frac{mu_m'(t)}{||u_m||_{1,4}}) - \frac{mu_m(t)}{||u_m||_{1,4}} = 0. \tag{29}$$

Multiply by $u_m(t)$ and integrate over $[0,1]$

$$\int_0^1 |u_m'(t)|^2 dt + \int_0^1 f(t, \frac{mu_m(t)}{|||u_m||_{1,4}}, \frac{mu_m'(t)}{||u_m||_{1,4}}) u_m(t) dt \le 0.$$

From condition (27) one gets

$$\int_0^1 |u_m'(t)|^2 dt + \int_0^1 \frac{||u_m||_{1,4}}{m} \{K(\frac{mu_m(t)}{||u_m||_{1,4}}) \times \frac{mu_m'(t)}{||u_m||_{1,4}} - \tag{30}$$

$$-\alpha \left| \frac{mu_m(t)}{||u_m||_{1,4}} \right|^2 + \gamma \} dt \le 0$$

that is

$$\int_0^1 |u_m'(t)|^2 dt - \alpha \int_0^1 |u_m(t)|^2 dt + \tag{31}$$

$$+ \frac{m^d}{||u_m||_{1,4}^d} \int_0^1 K(u_m(t)) \cdot u_m'(t) dt \le C.$$

Let H be the indefinite integral of K. Then $H(u_m(t))$ is the indefinite integral of $K(u_m(t)) \cdot u_m'(t)$ and from (28), (31) we infer

$$\int_0^1 |u_m'(t)|^2 dt - \alpha \int_0^1 |u_m(t)|^2 dt \leq C.$$

From the inequality

$$\int_0^1 |u_m(t)|^2 dt \leq \int_0^1 |u_m'(t)|^2 dt \tag{32}$$

it yields $\{u_m'\}$ to be bounded in $L^2(0,1)$, which combined with (28) gives $\{u_m\}$ bounded in $H^1(0,1)$ and in $C(0,1)$.

Now from (29) and (26) we get $\{u_m\}$ to be bounded in $W^{2,1}(0,1)$ that is, for instance, $\{u_m\}$ is bounded in $W^{1,4}(0,1)$ too.

So for a sufficiently large m we have $||u_m||_{1,4} \leq m$ and u_m verifies (24), (25) which finishes the proof.

Corollary 1 *Under the same hypotheses as Theorem 4, with $\alpha < 0$ in (27), the periodic problem*

$$-u_m''(t) + f(t, u(t), u'(t)) = 0 \quad \text{a.e. } [0,1]$$

$$u(0) = u(1),\ u'(0) = u'(1)$$

has at least one solution $u \in W^{2,1}(0,1)$.

The proof follows the same lines as in *Theorem 4* because the corresponding operators T and S, defined in this case, have the same properties and the estimations can be derived in a similar way.

Remark 5 The classical result of Bernstein [7] ensures the existence of a solution for the two point problem, provided $f(t,u,v)$, $\dfrac{\partial f}{\partial u}(t,u,v)$, $\dfrac{\partial f}{\partial v}(t,u,v)$ continuous on $(0,1) \times R \times R$ and

$$\frac{\partial f}{\partial u}(t,u,v) \geq K > 0 \tag{33}$$

and (26) with $g(t,u)$, $h(t,u)$ continuous in $(0,1) \times R$.

We use the Lagrange theorem

$$f(t,u,v) - f(t,0,v) = \frac{\partial f}{\partial u}(t,\tilde{u},v)\cdot u$$

where $\tilde{u}$ is some point between u and 0.
Multiply by u

$$f(t,u,v)\cdot u = \frac{\partial f}{\partial u}(t,\tilde{u},v)\cdot u^2 + f(t,0,v)\cdot u \geq Ku^2 + f(t,0,v)\cdot u$$

which may be more restrictive than (27).

Example 5 Let $f(t,u,v) = a(t)u + b(t)v + c(t)$. Then (33) requires $a(t) \geq K > 0$, while (27) with $K(u) = u$ is fulfilled when $a(t) \geq 0$ only.

Example 6 We give now an example when f has quadratic growth in v

$$f(t,u,v) = a(t)u^{2n+1} + b(t)u^p v + c(t)v^2 u + d(t).$$

Then $f(t,u,v)u \geq b(t)u^{p+1}\cdot v + d(t)\cdot u$ in case $a(t) \geq 0$, $c(t) \geq 0$ and (27) is fulfilled.
Condition (33) gives

$$(2n+1)a(t)u^{2n} + pb(t)u^{p-1}\cdot v + c(t)v^2 \geq K > 0$$

which fails for $u = v = 0$ for any $a(t), b(t), c(t)$.

6 Appendix

We give a result concerning the Nemitsky operator in Sobolev spaces. See also Marcus and Mizel [15] or Pascali and Sburlan [17], p. 165.

Let s, p, q be real numbers such that $W^{1,q}(\Omega) \subset L^p(\Omega)$ continuously i.e.

$$\frac{1}{s} \geq \frac{1}{q} - \frac{1}{N}.$$

Proposition 2 *Operator $S : W^{1,q}(\Omega) \to L^p(\Omega)$ defined by*

$$(Su)(x) = f(x, u(x), \text{grad } u(x))$$

where f satisfies the assumptions

$$f(x,\cdot,\cdot) \text{ is continuous a.e. } x \in \Omega \tag{34}$$

$$f(\cdot,u,v) \text{ is measurable for. every } u,v \tag{35}$$

$$|f(x,u,v)| \leq l(x) + h(x)|u|^{\frac{s}{p}} + K(x)|v|^{\frac{q}{p}} \tag{36}$$

with $l \in L^p(\Omega)$, $h, K \in L^\infty(\Omega)$, is bounded and continuous.

Proof

Using an argument with simple functions we see that S maps measurable functions in measurable functions. From condition (36) and $W^{1,q}(\Omega) \subset L^s(\Omega)$ continuously it yields that operator S is well-defined and bounded.

Consider now a sequence $\{u_n\} \subset W^{1,q}(\Omega)$ such that $u_n \to u$ in $W^{1,q}(\Omega)$, that is $u_n \to u$ in $L^s(\Omega)$ and grad $u_n \to$ grad u in $L^q(\Omega)$. To show that S is continuous it suffices to show that there is an infinite subsequence such that $S(u_j) \to S(u)$ strongly in $L^p(\Omega)$.

We choose an infinite subsequence of $\{u_n\}$, which we denote $\{u_j\}$, such that

$$\text{grad } u_j \to \text{grad } u \quad \text{a.e. } \Omega.$$

Then, by (34) $\quad S(u_j) \to S(u)$ a.e. in Ω.
From (36) it follows that functions $|f(x, u_j(x), \text{grad } u_j(x))|^p$ are equi-integrable over Ω, so the almost everywhere convergence of $S(u_j)$ to $S(u)$ implies that $S(u_j) \to S(u)$ strongly in $L^p(\Omega)$ and the proof is finished.

Corollary 2 *Operator $S : W^{1,q}(\Omega) \to L^p(\Omega)$ defined by*

$$(Su)(x) = g(x, u(x))$$

where g satisfies the assumptions

$$g(x,\cdot) \text{ is continuous a.e. } x \in \Omega \tag{37}$$

$$g(\cdot,u) \text{ is measurable for every } u \tag{38}$$

$$|g(x,u)| \leq l(x) + h(x)|u|^{\frac{S}{p}} \tag{39}$$

with $l \in L^p(\Omega)$, $h \in L^\infty(\Omega)$, is bounded and continuous.

Acknowledgments

This work was supported by Grant with contract 145/2011, CNCS Romania.

References

[1] J. P. Aubin. *Un théorème de compacité*. C. R. Acad. Sci. Paris, 256, (1963) 5042 - 5044.

[2] V. Barbu. *Nonlinear Semigroups and Differential Equations in Banach Spaces*. Noordhoff International Publishing, Leyden, the Netherlands - Bucureşti, (1976).

[3] H. Brezis. *Opérateurs maximaux monotones et sémigroupes de contractions dans les espaces de Hilbert*. Math. Studies, 5, North Holland, Amsterdam, (1973).

[4] H. Brezis. *Problémès unilateraux*. J. Math. Pures Appl., 51, (1972), 1 - 164.

[5] H. Brezis, A. Haraux. *Image d'une somme d'opérateurs monotones et applications*. Israel J. of Math., vol. 23, No. 2, (1976).

[6] H. Brezis, L. Nirenberg. *Characterizations of the Ranges of Some Nonlinear Operators and Applications to Boundary Value problems*. Ann. Sc. Norm. Sup. di Pisa, Vol. V, 2, (1978).

[7] S. N. Bernstein. *Sur les équations du calcul des variations*. Ann. Ec. Norm. Sup., 29 (1912), 431 - 486.

[8] W. Chikouche, D. Mercier, S. Nicaise. *Regularity of the solution of some unilateral boundary value problems in polygonal and polyhedral domains*. Comm. Part. Diff. Eq., Vol. 28, No. 11-12 (2003), p. 1475 - 2001.

[9] A. Domarkas. *Regularity of solution of quasilinear elliptic equations with unilateral boundary conditions*. Lithuanian Math. J., Vol. 20, No. 1 (1980), p. 8 - 13.

[10] G. Duvaut, J. L. Lions. *Sur les inéquations en mécanique et en physique*. Dunod, Paris, (1972).

[11] D. Goeleven. *Noncoercive variational problems and related results*. Longman, London (1996).

[12] D. Goeleven, D. Motreanu, Y. Dumont, M. Rochdi. *Variational and hemivariational inequalities*. Theory, Methods and Applications, Kluwer Academic Press, Norwell, MA (2003).

[13] N. Halidias. *Unilateral boundary value problems with jump discontinuities*. Int. J. Math. and Mathematical Sciences, Vol. 30 (2003), p. 1433 - 1941.

[14] Vy Khoi Le. *On some noncoercive variational inequalities containing degenerate elliptic operators*. ANZIAM J., Vol. 44 (2003), p. 409 - 430.

[15] M. Marcus, V. Mizel. *Nemitsky Operators on Sobolev Spaces*. Arch. Rat. Mech. Anal., Vol. 51, (1973), 347 - 370.

[16] J. P. Puel. *Inéquations d'evolution paraboliques avec convexes dépendent du temps. Applications aux inéquations quasi-variationelles d'evolution*. Arch. Rat. Mech. Anal., Vol. 64, No. 1, (1977).

[17] D. Pascali, S. Sburlan. *Nonlinear Mappings of Monotone Type*. Sijhtoff & Noordhoff International Publishers, Leyden, the Netherlands - Bucureşti, (1978).

[18] V. Rosalba. *Strong sovability of a unilateral boundary value problems for nonlinear parabolic operators*. Mediteranean J. of Math., Vol. 4, No. 1 (2007), p. 119 - 126.

[19] D. Tiba. *Nonlinear Boundary Value Problems for Second Order Differential Equations*. Funkcialaj Ecvacioj, 20, (1977).

[20] D. Tiba. *General Boundary Value Problems for Second Order Differential Equations*. Nonlinear Analysis, Theory, Methods & Applications, Vol. 2, No. 4, (1978), p. 447 - 455.

Ann. Acad. Rom. Sci.
Ser. Math. Appl.
ISSN 2066-6594 Vol. 7, No. 1/2015

A Survey of the P Function Method for Higher Order Equations and Some Applications*

Cristian - Paul Dăneţ†

Dedicated to the memory of Prof. Dr. Viorel Arnăutu

Abstract

This paper gives a survey of an extension of the classical maximum principle, namely the P function method.

MSC: 35B50, 35G15, 35J40.

keywords: maximum principle, P function method, higher order, elliptic, plate theory.

1 Introduction

The intention of this paper is to survey some extensions (the P function method) and applications of the classical maximum principle for elliptic operators.

It is well-known that every subharmonic function in a bounded domain Ω (i.e. $\Delta u \geq 0$ in Ω) satisfies the classical maximum principle

$$\max_{\overline{\Omega}} u = \max_{\partial\Omega} u.$$

*Accepted for publication in revised form on January 22-nd, 2015

†cristiandanet@yahoo.com Department of Applied Mathematics, University of Craiova, Al. I. Cuza St., 13, 200585 Craiova, Romania,

The subbiharmonic function $u(x) = -x_1^4 - |x|^2$ in the ball $\Omega = \{(x_1, \ldots, x_n) | |x| < R\}$ (i.e. $\Delta^2 u \leq 0$ in Ω) shows that there are no classical maximum principles for the biharmonic operator $\Delta^2 u$ (and for higher-order elliptic operators at all). Still some results can be proven.

The first proof of a maximum principle for an elliptic equation of higher-order that has a similar form to the classical maximum principle was given by Miranda [33].

Miranda showed that for the biharmonic equation $\Delta^2 u = 0$, where u is a smooth function defined on a plane domain the function $|\nabla u|^2 - u\Delta u$ takes its maximum value on the boundary of the domain. Later, in [37], Payne uses functionals containing the square of the second gradient of the solution to semilinear equations of the form

$$\Delta^2 u = f(u)$$

to deduce integral bounds on $(\Delta^2 u)^2$.

Since then many authors have extended the Miranda's result. For example, maximum principles for fourth order equations containing nonlinearities in u or Δu can be found in works of Payne [37], Schaefer [57], [60], [61]. Similar results are proved by Zhang [72], Mareno [30], [31] (studied some equations from plate theory), Danet [5], [6], [7], [9], Tseng and Lin [68], etc. (see the references cited here).

Most recently the authors in [9] obtain maximum principles results for the more general variable coefficient m- metaharmonic equation

$$\Delta^m u - a_{m-1}(x)\Delta^{m-1} u + a_{m-2}(x)\Delta^{m-2} u - \cdots + a_0(x)u = 0 \quad \text{in } \Omega. \quad (1)$$

using P functions containing terms of the form $(\Delta^i u)$. Here Ω is a bounded domain in $\mathbb{R}^n$.

The survey paper [8] is devoted to the P function method and gives a presentation of research of the past years on applications of the P function method in second order elliptic problems. Historical notes and an extensive survey of the literature is added. The present paper intends to continue our previous work [8] by presenting contributions to the P function method for higher order elliptic equations.

2 Main Results

2.1 The general case (m arbitrary)

First we present a maximum principle for the general equation of order $2m$.

Theorem 2.1 *([9]) Let u be a classical solution of (1), i.e. $C^{2m}(\Omega) \cap C^{2m-2}(\overline{\Omega})$. We consider the function* P_1

$$\mathrm{P}_1 = (\Delta^{m-1}u)^2 + 2a_{m-2}(\Delta^{m-2}u)^2 + (\Delta^{m-3}u)^2 + \cdots + u^2.$$

Suppose that a_{m-1}, $a_{m-2} > 0$ and $\Delta(1/a_{m-2}) \le 0$ in Ω. If

$$\sup_{\Omega}\left\{\frac{a_0^2}{2a_{m-1}+1}\right\} < \frac{4n+4}{(\operatorname{diam}\Omega)^2},$$

$$\frac{a_0^2}{2a_{m-1}} > \max\left\{1+\sup_{\Omega} a_1, \ldots, 1+\sup_{\Omega} a_{m-3}\right\},$$

$$\frac{a_0^2}{2a_{m-1}+1} > \sup\left\{|a_1| + \cdots + |a_{m-3}|\right\},$$

and

$$\left(\frac{a_0^2}{2a_{m-1}} + 1\right) a_{m-2} > 1 \quad \text{in } \Omega$$

then, either there exists a constant $k \in \mathbb{R}$ such that $\mathrm{P}_1/w_1 \equiv k$ in Ω or P_1/w_1 does not attain a nonnegative maximum in Ω. Here $w_1(x) = 1 - \alpha(x_1^2 + \cdots + x_n^2) \in C^\infty(\mathbb{R}^n)$, and α is a positive constant.

Remark. The coefficient a_0 can be replaced by a_{m-j}, $j = 4, \ldots, m-1$ if there exists a $j = 4, \ldots, m-1$ such that

$$\frac{a_{m-j}^2}{2a_{m-1}} > \max_{k=3,\ldots,m}\left\{2+\sup_{\Omega} a_k\right\},$$

$$\frac{a_{m-j}^2}{2a_{m-1}} + 2 > \sup_{\Omega}\left\{|a_0| + \cdots |a_{m-j-1}| + |a_{m-j+1}| + \cdots + |a_{m-3}|\right\},$$

and

$$\left(\frac{a_{m-j}^2}{2a_{m-2}} + 1\right) a_{m-2} > 1 \quad \text{in } \Omega.$$

We now show that the uniqueness result and the maximum principle holds ([9]).

Theorem 2.2 *There is at most one classical solution of the boundary value problem*

$$\begin{cases} \Delta^m u - a_{m-1}(x)\Delta^{m-1}u + a_{m-2}(x)\Delta^{m-2}u + \cdots + (-1)^m a_0(x)u = f & \text{in } \Omega \\ u = g_1,\ \Delta u = g_2, \ldots,\ \Delta^{m-1}u = g_m & \text{on } \partial\Omega, \end{cases} \tag{2}$$

provided the coefficients $a_{m-1}, \ldots, a_0$ satisfy the conditions imposed in Theorem 2.1.

Remark. The boundary value problem

$$\begin{cases} \Delta^m u + 2^m u = 0 & \text{in } \Omega = (0,\pi)\times(0,\pi) \\ u = \Delta u = \cdots = \Delta^{m-1}u = 0 & \text{on } \partial\Omega, \end{cases}$$

has (at least) the solutions $u_1(x,y) \equiv 0$ and $u_2(x,y) = \sin x \sin y$ in Ω. This example shows that if we do not impose some restrictions on the coefficients $a_{m-1}, \ldots, a_0$, then the uniqueness result might be violated.

In [2] the authors pose an interesting open problem: If $f = 0$ in $\Omega, g_2 = \cdots = g_m = 0$ on $\partial\Omega, m \geq 3, n \geq 2, a_{m-1} = \cdots = a_1 = 0$ in Ω do all the solutions of (2) satisfy the maximum principle (3) where $C > 1$ is a constant? This problem, as it turns out, can be solved when Ω is a class C^2 domain ([63]). Here we present a version for arbitrary domains ([9]).

Theorem 2.3 *We consider the boundary value problem (2), where $f = 0$ in Ω and $g_2 = \cdots = g_m = 0$ on $\partial\Omega$.*
Then

$$\max_{\overline{\Omega}} |u| \leq C \max_{\partial\Omega} |u|, \tag{3}$$

holds for all solutions of (2) provided the coefficients $a_{m-1}, \ldots, a_0$ are subject to one of the conditions imposed in theorem 2.1.

Theorem 2.4 *([32])*
Suppose that $u \in C^{2m+1}(\Omega) \cap C^{2m-1}(\overline{\Omega})$ is a solution of (1). Furthermore for $n > 4$ one defines

$$\mathrm{P}_2 = \nabla^2(\Delta^{m-2}u) \cdot \nabla^2(\Delta^{m-2}u) - \nabla(\Delta^{m-2}u) \cdot \nabla(\Delta^{m-1}u) +$$

$$+ \frac{a_{m-1}}{2}\nabla(\Delta^{m-2}u) \cdot \nabla(\Delta^{m-2}u) + \frac{a_{m-3}}{2}\nabla(\Delta^{m-3}u) \cdot \nabla(\Delta^{m-3}u)$$

$$+ a_{m-2}\left[\frac{n-4}{n+2}\right](\Delta^{m-2}u)^2 - \left[\frac{4-n}{2(n+2)}\right](\Delta^{m-1}u)^2 + \sum_{i=0}^{m-2}\phi_i(\Delta^i u)^2,$$

where the functions $\phi_0, \ldots, \phi_{m-2} \in C^0(\overline{\Omega}) \cap C^2(\Omega)$ *satisfy* $\sum_{i=0}^{m-2}\phi_i^2 + 1 \leq \alpha$ *for some positive constant* α, *and* $|\nabla^2 u|^2 = u_{,ij}\, u_{,ij}$. *Additionally, one imposes the conditions*

$$\sum_{i=0}^{m-3} a_i^2 \leq \beta,\ \sum_{i=0}^{m-1}\nabla a_i \cdot \nabla a_i \leq \gamma,$$

for constants $\gamma \geq 0$, $\beta > 0$.

$$\phi_i \geq \frac{\beta}{2}$$

$$a_{m-2} \geq 1,\ a_{m-1} - \frac{1}{2} \geq \frac{\gamma(n+2)}{2(n-4)},$$

$$\frac{\Delta a_i}{2} - \frac{\nabla a_{m-i}\cdot\nabla a_{m-i}}{a_{m-i}} \geq 0,\ i = 1, 3,$$

$$\Delta a_{m-2} - 4\frac{\nabla a_{m-2}\cdot\nabla a_{m-2}}{a_{m-2}} \geq 0,$$

$$\Delta\phi_i \geq 3\max\left\{\frac{\beta(n-4)}{2(n+2)} + \frac{\gamma}{2}, \alpha, 4\frac{\nabla\phi_i\cdot\nabla\phi_i}{\phi_i}\right\},\ i = 0, \ldots, m-2.$$

Then, P_2 *is subharmonic in* Ω.

We briefly indicate how theorem 2.4 can be used to obtain integral bounds on the square of the second gradient of $\Delta^{m-2}u$. Suppose that the hypotheses of theorem 2.4 are satisfied and the m conditions

$$\Delta^i u = 0,\ i = 0, \ldots, m-5,$$

$$\Delta^{m-2}u = \frac{\Delta^{m-2}u}{\partial n} = 0,\ \Delta^{m-3}u = \frac{\Delta^{m-3}u}{\partial n} = 0,$$

hold on $\partial\Omega$. Let A denote the area of Ω. Using theorem 2.4 and integration by parts we get

$$\int_\Omega |\nabla^2(\Delta^{m-2}u)|^2dx \leq \frac{A}{2}\max_{\partial\Omega}\Bigg\{|\nabla^2(\Delta^{m-2}u)|^2+\phi_{m-4}(x)(\Delta^{m-4}u)^2+$$

$$\frac{n-4}{2(n+2)}(\Delta((\Delta^{m-2}u)))^2\Bigg\}.$$

Before treating some particular cases, we shift our attention from n dimensional to one dimensional case and mention the following result (Ω denotes an open interval (α,β)) ([5])

Theorem 2.5 *There can be at most one classical solution of the problem*

$$\begin{cases} u^{(2m)}-du^{(6)}+c(x)u^{(4)}-b(x)u''+a(x)u=f & \text{in } \Omega \\ u=g_1,\ u''=g_2,\ u'''=g_3,\dots,u^{(m)}=g_m & \text{on } \partial\Omega, \end{cases}$$

where $m\geq 6$ is even, $d\geq 0$ and $b\geq 0$, $a,c>0$, $(1/a)'',(1/c)''\leq 0$ in Ω.

The result follows since the function

$$\begin{aligned} \mathrm{P}_3 \;=\;& u''u^{(2m-2)}-2u'''u^{(2m-3)}+3u^{(4)}u^{(2m-4)}-\dots+(m-3)u^{(m-2)}u^{(m+2)} \\ & -(m-3)u^{(m-1)}u^{(m+1)}/2-((m-3)/2+1)[\,(u^{(m)})^2-u^{(m-1)}u^{(m+1)}] \\ & +[(u^{'''})^2-du^{''}u^{(4)}]+c(x)(u'')^2/2+a(x)u^2/2 \end{aligned}$$

assumes its maximum value on $\partial\Omega$, where u is a solution of

$$u^{(2m)}-du^{(6)}+c(x)u^{(4)}-b(x)u''+a(x)u=0 \quad \text{in } \Omega.$$

Similarly, we can treat the problem

$$\begin{cases} u^{(2m)}+du^{(6)}-c(x)u^{(4)}+b(x)u''-a(x)u=f & \text{in } \Omega \\ u=g_1,\ u''=g_2,\ u'''=g_3,\dots,u^{(m)}=g_m & \text{on } \partial\Omega, \end{cases}$$

where $m\geq 5$ is odd.

2.2 The particular case $m = 4$

In this section we consider classical solutions (i.e., $C^8(\Omega) \cap C^6(\overline{\Omega})$) of

$$\Delta^4 u - a(x)\Delta^3 u + b(x)\Delta^2 u - c(x)\Delta u + du = 0, \tag{4}$$

in the bounded plane domain Ω, and present ([5]) a maximum principle for a certain function defined on the solutions of (4). Then we use the maximum principle to prove a uniqueness result for the corresponding boundary value problem.

Theorem 2.6 *Let u be a classical solution of (4). Assume that*

$$a > 0, \quad \Delta(1/a) \leq 0 \quad \text{in } \Omega,$$

$$b \geq 0 \quad \text{in } \Omega,$$

$$c > 0, \quad \Delta(1/c) \leq 0 \quad \text{in } \Omega,$$

and

$$d > 0$$

are satisfied. Then the functional

$$\mathrm{P}_4 = \frac{c(x)}{2}(\Delta u)^2 + \frac{a(x)}{2}(\Delta^2 u)^2 + d(|\nabla u|^2 - u\Delta u) + |\nabla(\Delta^2 u)|^2 - \Delta^2 u \Delta^3 u$$

assumes its maximum value on $\partial\Omega$. The result also holds if a and c are nonnegative constants.

An important application of the above presented maximum principle is the following uniqueness result:

Theorem 2.7 *There is at most one classical solution of the boundary value problem*

$$\begin{cases} \Delta^4 u - a\Delta^3 u + b(x)\Delta^2 u - c\Delta u + du = f & \text{in } \Omega, \\ u = g, \quad \Delta u = h, \quad \Delta^2 u = i, \quad \Delta^3 u = j & \text{on } \partial\Omega, \end{cases} \tag{5}$$

where $a, c \geq 0$, b and d satisfy the hypotheses of theorem 2.6, and the curvature k of $\partial\Omega$ (Ω is a smooth plane domain) is strictly positive.

We suppose that u_1 and u_2 are two solutions of (5). Defining $v = u_1 - u_2$, we see that v satisfies (4) and

$$v = \Delta v = \Delta^2 v = \Delta^3 v = 0 \quad \text{on } \partial\Omega. \tag{6}$$

By virtue of theorem 2.6

$$\mathrm{P}_4 \leq \max_{\partial\Omega} \mathrm{P}_4 \quad \text{in } \Omega. \tag{7}$$

Since $v = \Delta^2 v = 0$ on $\partial\Omega$, we have

$$|\nabla v| = \left|\frac{\partial v}{\partial n}\right| \quad \text{on } \partial\Omega \tag{8}$$

and

$$|\nabla(\Delta^2 v)| = \left|\frac{\partial(\Delta^2 v)}{\partial n}\right| \quad \text{on } \partial\Omega, \tag{9}$$

where $\partial/\partial n$ denotes the outward directed normal derivative operator.
It can be shown that (introducing a normal coordinate system)

$$\frac{\partial v}{\partial n} = \frac{\partial(\Delta^2 v)}{\partial n} = 0 \quad \text{on } \partial\Omega. \tag{10}$$

By (6), (7), (8), (9) and (10) we get

$$\mathrm{P}_4 \leq 0 \quad \text{in } \Omega,$$

which gives

$$-v\Delta v - \Delta^2 v \Delta^3 v \leq 0 \quad \text{in } \Omega. \tag{11}$$

Integrating (11) over Ω and using Green's identity we obtain

$$\int_\Omega |\nabla v|^2 + \int_\Omega |\nabla(\Delta^2 v)|^2 \leq 0.$$

Hence $v \equiv 0$ in $\overline{\Omega}$ by continuity.

It is known that once we have a maximum principle for an equation, the nonexistence of a nontrivial solution of the zero-boundary problem will be a consequence.
An inverse result, of establishing a maximum principle from some nonexistence results was carried out by Schaefer and Walter (Theorem 2, [63]).
Using their result and our theorem 2.7, we obtain the following maximum principle

Corollary 2.1 *Suppose that u is a classical solution of the boundary value problem*

$$\begin{cases} \Delta^4 u - a\Delta^3 u + b\Delta^2 u - c\Delta u + du = 0 & \text{in } \Omega, \\ \Delta u = 0, \quad \Delta^2 u = 0, \quad \Delta^3 u = 0 & \text{on } \partial\Omega, \end{cases}$$

where $a, b, c \geq 0$, d satisfy the hypotheses of theorem 2.6, and the curvature k of $\partial\Omega$ (Ω is a smooth domain) is strictly positive. Then there exists a constant $K > 0$ such that

$$\max_{\overline{\Omega}} |u| \leq K \max_{\partial\Omega} |u|.$$

Remark.

1. Similar uniqueness results can be inferred using theorem 2.6. It can be shown (see [5]) that there is at most one classical solution of the boundary value problem

$$\begin{cases} \Delta^4 u - a\Delta^3 u + b(x)\Delta^2 u - c\Delta u + du = f & \text{in } \Omega \\ u = g, \quad \Delta u = h, \quad \Delta^2 u = i, \quad \frac{\partial(\Delta^2 u)}{\partial n} = j & \text{on } \partial\Omega, \end{cases}$$

2. We note that Dunninger [11] developed a maximum principle from which follows the uniqueness for the classical solution of the boundary value problem

$$\begin{cases} \Delta^2 u + cu = f & \text{in } \Omega \subset \mathbb{R}^n, \\ u = g, \Delta u = h & \text{on } \partial\Omega, \end{cases}$$

where $c > 0$ is a constant.

An uniqueness result for solutions of a more general fourth-order elliptic equation, under the same boundary conditions follows from Corollary 1, [72].

The uniqueness question for solutions of the boundary value problem (here $a, b \geq 0$ and $c > 0$ in Ω)

$$\begin{cases} \Delta^3 u - a(x)\Delta^2 u + b(x)\Delta u - c(x)u = f & \text{in } \Omega \subset \mathbb{R}^n, \\ u = g, \ \Delta u = h, \ \Delta^2 u = i & \text{on } \partial\Omega, \end{cases}$$

has been settled in a satisfactory way by Schaefer [58] (the constant coefficient case with n=2) and Goyal and Goyal [17] (the constant and variable coefficient case).

We see that our uniqueness result (theorem 2.7) is a generalization of results of Dunninger, Goyal and Schaefer.

2.3 The particular case $m = 3$.

This subsection is dedicated to maximum principles for a class of linear equations of sixth order. As a consequence of these maximum principles we will obtain uniqueness results for boundary value problems of sixth order. This section is based on the paper [7].

Schaefer [58] investigated the uniqueness of the solution for the boundary value problems

$$\begin{cases} \Delta^3 u - a(x)\Delta^2 u + b(x)\Delta u - c(x)u = f & \text{in } \Omega \subset \mathbb{R}^n \\ u = g,\ \Delta u = h,\ \Delta^2 u = i & \text{on } \partial\Omega, \end{cases} \tag{12}$$

where $a, b, \geq 0$, $c > 0$ are constants, and the curvature of $\partial\Omega$ is positive.

Our aim here is to remove via the P function method dimension and geometry conditions (convexity and smoothness) with, of course, further conditions on the coefficients a, b and c.

We deal with classical solutions (i.e. $u \in C^6(\Omega) \cap C^4(\overline{\Omega})$) of

$$\Delta^3 u - a(x)\Delta^2 u + b(x)\Delta u - c(x)u = 0 \quad \text{in } \Omega \subset \mathbb{R}^n, n \geq 2. \tag{13}$$

The uniqueness results can be inferred from the following maximum principles.

Theorem 2.8 *Let u be a classical solution of (13) and suppose that*

$$\frac{a(b+c)^2}{b^2(a-1)} < \frac{8n+8}{(\operatorname{diam}\Omega)^2}, \tag{14}$$

holds, where $a > 1, b, c$ are constants. We consider the function P_5 given by

$$\mathrm{P}_5 = (a\Delta^2 u + bu)^2 + ab(a-1)(\Delta u)^2 + b^2(a-1)u^2.$$

Then, either there exists a constant $k \in \mathbb{R}$ such that $\mathrm{P}_5/w_1 \equiv k$ in Ω or P_5/w_1 does not attain a nonnegative maximum in Ω.

By computation and using equation (13) we have in Ω

$$\begin{aligned} \Delta\big((a\Delta^2 u + bu)^2\big) &\geq 2(a\Delta^2 u + bu)(a\Delta^3 u + b\Delta u) \\ &= 2\big(a^3(\Delta^2 u)^2 + abcu^2 + a^2(b+c)u\Delta^2 u + \\ &\quad ab(1-a)\Delta u\Delta^2 u + b^2(1-a)u\Delta u\big), \end{aligned}$$

$$\Delta\big(ab(a-1)(\Delta u)^2\big) \geq 2ab(a-1)\Delta u\Delta^2 u,$$

$$\Delta\big(b^2(a-1)u^2\big) \geq 2b^2(a-1)u\Delta u.$$

That means that

$$\begin{aligned}\Delta \mathrm{P}_5 &\geq 2a\big(a^2(\Delta^2 u)^2 u + a(b+c)u\Delta^2 u + bcu^2\big)\\ &= 2a\left(a\Delta^2 u + \frac{b+c}{2}u\right)^2 + 2a\left(bc - \frac{(b+c)^2}{4}\right)u^2 \qquad (15)\\ &\geq -\frac{a(b+c)^2}{2}u^2.\end{aligned}$$

Hence P_5 satisfies the differential inequality

$$\Delta \mathrm{P}_5 + \frac{a(b+c)^2}{2b^2(a-1)}\mathrm{P}_5 \geq 0 \quad \text{in } \Omega.$$

Since (14) holds, we can use a version of the generalized maximum principle (lemma 2.1, [7]) to obtain the desired result.

Theorem 2.9 *Let u be a classical solution of (13) and suppose that*

$$\sup_{\Omega} \frac{(a+c)^2}{a(b-1)} < \frac{8n+8}{(\operatorname{diam}\Omega)^2}, \qquad (16)$$

$$b > 1 \quad \text{in } \overline{\Omega},\ \Delta(1/(\mathrm{b}-1)) \leq 0 \quad \text{in } \Omega$$

holds.

If

$$\mathrm{P}_6 = (\Delta^2 u + u)^2 + (b-1)(\Delta u)^2 + (b-1)u^2$$

then, either there exists a constant $k \in \mathbb{R}$ such that $\mathrm{P}_6/w_1 \equiv k$ in Ω or P_6/w_1 does not attain a nonnegative maximum in Ω.

If $a = c$ in Ω then, P_6 attains its maximum value on $\partial\Omega$ (the restriction (16) is not needed).

Theorem 2.10 *Let u be a classical solution of (13), where $a > 0$ in $\overline{\Omega}$, and c is of arbitrary sign in Ω. Suppose that*

$$\sup_{\Omega} \frac{c^2}{2a} + 1 < \frac{4n+4}{(\operatorname{diam}\Omega)^2},$$

$$b > 0,\ \Delta(1/b) \leq 0 \quad \text{in } \Omega,$$

$$b\left(\frac{c^2}{2a}+1\right)\geq 1 \quad \text{in } \Omega$$

holds.

If

$$\mathrm{P}_7=(\Delta^2 u)^2+b(\Delta u)^2+u^2$$

then, either there exists a constant $k\in\mathbb{R}$ *such that* $\mathrm{P}_7/w_1\equiv k$ *in* Ω *or* P_7/w_1 *does not attain a nonnegative maximum in* Ω*.*

Theorem 2.11 *Let* u *be a classical solution of (13) and suppose that*

$$\sup_{\Omega}\frac{1}{a}\left(c+\frac{(c+1)^2}{4(a-1)}\right)<\frac{2n+2}{(\operatorname{diam}\Omega)^2},$$

$$b=0,\ a>1 \quad \text{in } \overline{\Omega},\ \Delta(1/\mathrm{a})\leq 0 \quad \text{in } \Omega,$$

$$c>0,\ \Delta(1/c)\leq 0 \quad \text{in } \Omega$$

holds.

We consider the function P_8 *given by*

$$\mathrm{P}_8=(\Delta^2 u-\Delta u)^2+c(\Delta u-u)^2+a(\Delta u)^2.$$

Then, either there exists a constant $k\in\mathbb{R}$ *such that* $\mathrm{P}_8/w_1\equiv k$ *in* Ω *or* P_8/w_1 *does not attain a nonnegative maximum in* Ω*.*

Now an uniqueness result follows from the above mentioned maximum principles.

Theorem 2.12 *There is at most one classical solution of the boundary value problem (12), where* a, b *and* c *satisfy the conditions of Theorem 2.8 or Theorem 2.9 or Theorem 2.10 or Theorem 2.11.*

For various uniqueness results for sixth order boundary value problems the reader is referred to [7].

2.4 The particular case $m = 2$.

In 1971, J. Serrin [65] and H. Weinberger [71] proved that if Ω is a bounded domain in $\mathbb{R}^n$ with smooth boundary

$$\begin{cases} \Delta u = -1 & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \\ \frac{\partial u}{\partial n} = c & \text{on } \partial\Omega, \end{cases}$$

(where c is a constant) then Ω is a ball of radius $|nc|$ and the solution is radially symmetric about the center.
Serrin's proof is based on the classical maximum principle and on the method of moving parallel planes.Weinberger's method is more elementary. It also uses the maximum principle but relies on Green's theorem to establish certain identities. Unfortunately, Weinberger's argument does not extend to more general results.

Using the following maximum principle Benett [1] was able to show that an analogous result holds for a fourth order problem.

Theorem 2.13 *The function*

$$\mathrm{P}_9 = \sum_{i,j=1}^{n} \frac{\partial^2 u}{\partial u_i \partial u_j} - \nabla u \cdot \nabla(\Delta u) + \frac{n-4}{n+2} \int_0^u f(y)dy + \frac{n-4}{2(n+2)} (\Delta u)^2$$

assumes its maximum value on $\partial\Omega$, where u is a solution of $\Delta^2 u + f(u) = 0$ in $\Omega \subset \mathbb{R}^n$, $f' \leq 0$ in $\mathbb{R}$.

Corollary 2.2 *([1]) Let Ω be a bounded domain in $\mathbb{R}^n$ with $C^{4+\varepsilon}$ boundary, and suppose that the following overdetermined problem has a solution in $u \in C^4(\partial\Omega)$*

$$\begin{cases} \Delta^2 u = -1 & \text{in } \Omega, \\ \frac{\partial u}{\partial n} = 0 & \text{on } \partial\Omega, \\ \Delta u \equiv c & \text{on } \partial\Omega \ (c \text{ - constant}). \end{cases}$$

Then Ω is an open ball of radius $[|c|(n^2+2n)]^{\frac{1}{2}}$ and u is radially symetric.

The above mentioned result allows a characterization of open balls in $\mathbb{R}^n$ by means of an integral identity:
Let Ω be a smooth bounded domain in $\mathbb{R}^n$ and suppose that there is a real constant M so that

$$\int_\Omega B dx = M \int_{\partial\Omega} \frac{\partial B}{\partial n} ds$$

holds for any function B in $C^4(\overline{\Omega})$ satisfying

$$\begin{cases} \Delta^2 B = 0 & \text{in } \Omega, \\ B = 0 & \text{on } \partial\Omega. \end{cases}$$

Then Ω is an open ball.

Finally, we state our last maximum principle for an fourth order equation.

Theorem 2.14 *([7])*
Let u be a classical solution of

$$\Delta^2 u - a_1 \Delta u + a_0(x) u = 0 \quad \text{in } \Omega \subset \mathbb{R}^n, \tag{17}$$

where $a_1 \equiv const. > 0$, $a_0 > 0$ in $\overline{\Omega}$.
Suppose that

$$\sup_{\Omega} \left(a_1 - \frac{1}{a_1} \left(\frac{a_0 - 1}{a_0} \right)^2 \right) < \frac{2n+2}{(\operatorname{diam} \Omega)^2}. \tag{18}$$

Let

$$\mathrm{P}_{10} = \frac{1}{2}(\Delta u - a u)^2 + \frac{1}{2}(\Delta u)^2 + u^2.$$

Then, either there exists a constant $k \in \mathbb{R}$ such that $\mathrm{P}_{10}/w_1 \equiv k$ in Ω or P_{10}/w_1 does not attain a nonnegative maximum in Ω.
If

$$a_1^2 \geq \left(\frac{a_0 - 1}{a_0} \right)^2 \quad \text{in } \Omega, \tag{19}$$

then the function P_{10} attains its maximum value on $\partial\Omega$ (here the assumption (18) is not needed).

Remark. A classical result ([2]) tells us that the boundary value problem

$$\begin{cases} \Delta^2 u - a_1(x) \Delta u + a_0(x) u = f & \text{in } \Omega \subset \mathbb{R}^n \\ u = g, \ \Delta u = h & \text{on } \partial\Omega, \end{cases} \tag{20}$$

has a unique solution if $a_1, a_0 > 0$ and if $\Delta a_0 < 0$ or $\Delta(1/a_0) < 0$ in Ω.
Theorem 2.14 tells us that if $a_1 \geq 1$ and $a_0 > 0$ then the boundary value problem (20) has a unique solution. We see that no smoothness restrictions are needed on the coefficient a_0.

References

[1] A. Bennett, Symmetry in an overdetermined fourth order elliptic boundary value problem, Siam J. Math. Anal. 17, (1986), 1354-1358.

[2] S. N. Chow, D. R. Dunninger, A maximum principle for n-metaharmonic functions, Proc. Amer. Math. Soc. 43 (1974), 79–83.

[3] S. N. Chow, D. R. Dunninger, A maximum principle for n-metaharmonic functions, Proc. Amer. Math. Soc. 43, (1974), 79-83.

[4] C. Cosner, P. W. Schaefer, On the development of functionals which satisfy a maximum principle, Appl. Anal. 26, (1987) 45-60.

[5] C. - P. Danet, Uniqueness results for a class of higher - order boundary value problems, Glasgow Math. J. 48, (2006), 547–552.

[6] C. - P. Danet, On the elliptic inequality L$u \leq 0$, Math. Inequal. and Appl. 11, (2008), 559–562.

[7] C. -P. Danet, Uniqueness in some higher order elliptic boundary value problems in n dimensional domains, Electronic J. Qualitative Theory of Diff. Eq., 54 (2011) 1–12.

[8] C. -P. Danet, The Classical Maximum Principle. Some of Its Extensions and Applications, Annals of the Academy of Romanian Scientists Series on Mathematics and its Applications, 3, No. 2 (2011), 273–299.

[9] C. -P. Danet and A. Mareno, Maximum principles for a class of linear equations of even order, Math. Inequal. Appl. 16 (2013), 809- 822

[10] C. -P. Danet, Two maximum principles for a nonlinear fourth order equation from thin plate theory, Electronic J. Qualitative Theory of Diff. Eq. 31 (2014), 1–9.

[11] D. R. Dunninger, Maximum principles for solutions of some fourth - order elliptic equations, J. Math. Anal. Appl. 37 (1972), 655–658.

[12] D. R. Dunninger, Maximum principles for fourth order ordinary differential inequalities, J. Math. Anal. Appl. 82, (1981), 399-405.

[13] R. J. Duffin, On a question of Hadamard concerning super-biharmonic functions, J. Math. Phys. 27, (1949), 253-258.

[14] C. F. Gauss, Allgemeine Theorie des Erdmagnetismus. Beobachtungen des magnetischen Vereins im Jahre 1838, Leipzig, 1839.

[15] S. Goyal and V. B. Goyal, Liouville - type and uniqueness results for a class of sixth-order elliptic equations, J. Math. Anal. Appl., 139, (1989), 586–599.

[16] V. B. Goyal, Liouville - type results for fourth order elliptic equations, Proc. Roy. Soc. Edinburgh, 103A, (1986), 209-213.

[17] S. Goyal, V. B. Goyal, Liouville - type and uniqueness results for a class of sixth - order elliptic equations, J. Math. Anal. Appl., 139, (1989), 586-599.

[18] S. Goyal, V. B. Goyal, Liouville - type and uniqueness results for a class of elliptic equations, J. Math. Anal. Appl. 151, (1990) 405-416.

[19] V. B. Goyal, P. W. Schaefer, Liouville theorems for elliptic systems and nonlinear equations of fourth order, Proc. Roy. Soc. Edinburgh, 91A, (1982), 235-242.

[20] V. B. Goyal, P. W. Schaefer, Comparison principles for some fourth order elliptic problems, in "Lecture Notes in Mathematics, 964, 272-279, Springer, 1982.

[21] V. B. Goyal, K. P. Singh, Maximum principles for a certain class of semi - linear elliptic partial differential equations, J. Math. Anal. Appl. 69, (1979), 1-7.

[22] D. Gilbarg, N. S. Trudinger, Elliptic Partial Differential Equations of Second Order, Classics in Mathematics, Springer, 2001.

[23] J. Hadamard, Mémoire sur le problème d'analyse relatif à l'équilibre des plaques élastiques encastrées, in: Oeuvres de Jacques Hadamard, Tome II, pp. 515-641, Centre National de la Recherche Scientifique: Paris, 1968, Reprint of: Mémoires présentés par divers savants à l'Académie des Sciences 33, (1908), 1-128.

[24] Z. Hailiang, Maximum principles for a class of semilinear elliptic boundary - value problems, Bull. Austral. Math. Soc. 52, (1995), 169-176.

[25] G. N. Hille, C. Zhou, Liouville theorems for even order elliptic systems, Indiana Univ. Math. J. 43, (1994) 383-410.

[26] E. Hopf, Elementare Bemerkungen über die Lösungen partieller Differential gleichungen zweiter Ordnung vom elliptischen Typus, Sitz. Ber. Preuss. Akad. Wissensch. Berlin, Math.-Phys. Kl 19, (1927), 147-152.

[27] E. Hopf, A remark on linear elliptic differential equations of second order, Proc. Amer. Math. Soc. 3, (1952), 791-793.

[28] A. Henrot and G. A. Philippin, Some overdetermined boundary value problems with elliptical free boundaries, SIAM J. Math. Anal. 28, (1998), 309-320.

[29] L. G. Makar-Limanov, Solutions of Dirichlet problem for the equation $\Delta u = -1$ in a convex region, math. Notes Acad. Sci. URSS, 9, (1971), 52-53.

[30] A. Mareno, Maximum principles for a fourth order equation from plate theory, J. math. Anal. Appl. 343 (2008), 932-937.

[31] A. Mareno, Integral Bounds for von Kármán's equations, Z. Angew. Math. Mech. 90, (2010), 509-513.

[32] A. Mareno, On Maximum Principles for -Metaharmonic Equations, ISRN Mathematical Analysis, vol. 2012, Article ID 634316, 9 pages, 2012. doi:10.5402/2012/634316

[33] C. Miranda, Formule di maggiorazione e teorema di esistenza per le funzioni biarmoniche di due variabili, Giorn. Mat. Battaglini 78, 97-118, 1948.

[34] C. Miranda, Teorema del massimo modulo e teorema di esistenza e di unicita per il problema di Dirichlet relative alle equazioni ellitiche in due variabili, Ann. Mat. Pura Appl. 46, (1958), 265-311.

[35] M. A. Paraf, Sur le problème de Dirichlet et son extension au cas de l'équation linéare générale du second ordre, Ann. Fac. Sci. Toulouse 6 (1982), Fasc. 47-Fasc. 54.

[36] L. E. Payne, Bounds for the maximal stressin the Saint Venant torsion problem, Indian J. Mech. Math. special issue, (1968), 51–59.

[37] L. E. Payne, Some remarks on maximum principles, J. Analyse Math. 30, (1976), 421-433.

[38] L. E. Payne, G. A. Philippin, Some applications of the maximum principle in the problem of torsional creep, Siam. J. Appl. Math. 33, (1977), 446-455.

[39] L. E. Payne, G. A. Philippin, Some maximum principles for nonlinear elliptic equations in divergence form with applications to capillary surfaces and to surfaces of constant mean curvature, Nonlinear Analysis 3, (1979), 193-211.

[40] L. E. Payne, G. A. Philippin, On maximum principles for a class of nonlinear second order elliptic equations, J. Diff. Eq. 37, (1980), 39-48.

[41] L. E. Payne, G. A. Philippin, On gradient maximum principles for quasilinear elliptic equations, Nonlinear Analysis 23, (1994), 387-398.

[42] L. E. Payne, G. A. Philippin, On some maximum principles involving harmonic functions and their derivatives, SIAM J. Math. Anal. 10, (1979), 96-104.

[43] L. E. Payne, P. W. Schaefer On overdetermined boundary value problems for the biharmonic operato, J. Math. Anal. Appl. 187, (1994), 598-616.

[44] L. E. Payne, R. P. Sperb, I. Stakgold, On Hopf type maximum principles for convex domains, Nonlinear Analysis 1, (1977), 547-559, 1977.

[45] L. E. Payne, I. Stakgold, Nonlinear problems in nuclear reactor analysis, Springer Lecture Notes in Mathematics, 322, 1973.

[46] L.E. Payne, P.W. Schaefer, Eigenvalue and eigenfunction inequalities for the elastically supported membrane, Z. angew. Math. Phys. 52 (2001), 888-895.

[47] L.E. Payne and G. A. Philippin, Isoperimetric inequalities in the torsion and clamped membrane problems for convex plane regions, SIAM J. Math. Anal. 14, (1983), 1154-1162.

[48] L.E. Payne and H.F. Weinberger, Lower bounds for vibration frequencies of elastically supported membranes annd plates, J.Soc. Ind. Appl. Math. 5 (1957), 171-182.

[49] G. Philippin, Some remarks on the elastically suported membrane, Z.A.M.P. 29, (1978), 306-314.

[50] G. A. Philippin, A. Safoui, Some minimum principles for a class of elliptic boundary value problems, Appl. Anal., (2004), 231-241.

[51] G. A. Philippin, A. Safoui, Some applications of the maximum principle to a variety of fully nonlinear elliptic PDE's, Z.A.M.P. 54, (2003), 739-755.

[52] M. H. Protter, H. F. Weinberger, Maximum Principles in Differential Equations, Prentice Hall Inc., 1967.

[53] G.Porru, A. Safoui and S. Vernier-Piro, Best possible maximum principles for fully nonlinear elliptic partial differential equations, J. Anal. and Appl. 25, (2006), 421-434.

[54] P. Pucci, J. Serrin, H. Zou, A strong maximum principle and a compact support principle for singular elliptic inequalities, J. Math. Pures Appl. 78, (1999), 769-789.

[55] M. H. Protter and H. F. Weinberger, Maximum principles in differential equations, Prentice - Hall, Inc., Englewood Cliffs, N. J., 1967.

[56] P. Quittner and P. Souplet, Superlinear parabolic problems, Birkhäuser, 2007.

[57] P. W. Schaefer, On a maximum principle for a class of fourth - order semilinear elliptic equations, Proc. Roy. Soc. Edinburgh Sect. A 77, (1977), 319-323.

[58] P. W. Schaefer, Uniqueness in some higher order elliptic boundary value problems, Z. Angew. Math. Mech. 29, (1978), 693-697.

[59] P. W. Schaefer, Some maximum principles in semilinear elliptic equations, Proc. Amer. Math. Soc. 98, (1986), 97-102.

[60] P. W. Schaefer, Pointwise estimates in a class of fourth - order nonlinear elliptic equations, Z. A. M. P. 38, (1987), 477-479.

[61] P. W. Schaefer, Solution, gradient, and laplacian bounds in some nonlinear fourth order elliptic equations, Siam J. Math. Anal. 18, (1987), 430-434.

[62] P. W. Schaefer (Editor), Maximum principles and eigenvalue problems in partial differential equations, Longman Scientific and Technical, 1988.

[63] P. W. Schaefer, W. Walter, On pointwise estimates for metaharmonic functions, J. Math. Anal. Appl. 69, 171-179, 1979.

[64] J. Serrin, The problem of Dirichlet for quasilinear elliptic differential equations with many independent variables, Philos. Trans. Roy. Soc. London Ser. A 264, (1969), 413-496.

[65] J. B. Serrin, A symmetry problem in potential theory, Arch. Rat. Mech. Anal. 43, (1971), 304-318.

[66] R. P. Sperb, Untere un obere Schranken für den tiefsten Eigenwert der elastisch gestützen Membran, Z. angew. Math. Phys. 23, (1972), 231-244.

[67] R. P. Sperb, Maximum Principles and Their Applications, Academic Press, 1981.

[68] S. Tseng, C-S. Lin, On a subharmonic functional of some even order elliptic problems, J. Math. Anal. Appl. 207, (1997), 127- 157.

[69] J. Urbas, Nonlinear oblique boundary value problem for Hessian equations in two dimensions, Ann. Inst. Henri Poincaré, 12, (1995) 507-575.

[70] J. R. L. Webb, Maximum principles for functionals associated with the solution of semilinear elliptic boundary value problems, Z.A.M.P. 40, (1989), 330-338.

[71] H. F. Weinberger, Remark on a preceding paper of Serrin, Arch. Rat. Mech. Anal. 43 (1971), 319-320.

[72] H. Zhang and W. Zhang, Maximium principles and bounds in a class of fourth - order uniformly elliptic equations, J. Phys. A : Math. Gen., 35, (2002), 9245-9250.

Ann. Acad. Rom. Sci.
Ser. Math. Appl.
ISSN 2066-6594 Vol. 7, No. 1/2015

A posteriori error identities for nonlinear variational problems*

Pekka Neittaanmaki† Sergey Repin‡

Dedicated to the memory of Prof. Dr. Viorel Arnăutu

Abstract

A posteriori error estimation methods are usually developed in the context of upper and lower bounds of errors. In this paper, we are concerned with a posteriori analysis in terms of identities, i.e., we deduce error relations, which holds as equalities. We discuss a general form of error identities for a wide class of convex variational problems. The left hand sides of these identities can be considered as certain measures of errors (expressed in terms of primal/dual solutions and respective approximations) while the right hand sides contain only known approximations. Finally, we consider several examples and show that in some simple cases these identities lead to generalized forms of the Prager-Synge and Mikhlin's error relations. Also, we discuss particular cases related to power growth functionals and to the generalized Stokes problem.

MSC: 65F30, 65F50, 65N35, 65F10

keywords: Estimates of deviations from the exact solution, convex variational problems, error measures for nonlinear problems.

*Accepted for publication on January 23-rd, 2015

†pn@mit.jyu.fi University of Jyväskylä, FIN–40014, P.O. Box 35 (Agora), Jyväskylä, Finland

‡serepin@jyu.fi University of Jyväskylä, FIN–40014, P.O. Box 35 (Agora), Jyväskylä and Saint Petersburg Polytechnic University, Polytechnicheskaya 29, 195251, St. Petersburg, Russian Federation

1 Functional setting

In this section, we present the class of variational problems to be considered and recall several basic facts related to this class of problems.

Throughout the paper we use two pairs of mutually conjugate reflexive Banach spaces. The first pair is Y and Y^* (with the duality pairing (y^*, y), where $y^* \in Y^*$ and $y \in Y$). The norms of Y and Y^* are denoted by $\|\cdot\|$ and $\|\cdot\|_*$, respectively. Another pair of spaces is V and V^*. The product of $v \in V$ and $v^* \in V^*$ is denoted by $< v^*, v >$. We assume that

$$V \subset \mathcal{V} \subset V^*,$$

where $\mathcal{V}$ is a Hilbert space with the norm $\|\cdot\|_{\mathcal{V}}$ and scalar product $(\cdot,\cdot)_{\mathcal{V}}$, so that $< v^*, v > = (v^*, v)_{\mathcal{V}}$ for any $v^* \in \mathcal{V}$.

By $\Lambda : V \to Y$ we denote a bounded linear operator and assume that the conjugate operator $\Lambda^* : Y^* \to V^*$ satisfies the relation

$$(y^*, \Lambda w) = < \Lambda^* y^*, w >, \qquad \forall w \in V. \tag{1.1}$$

If y^* is more regular and belongs to the set

$$H^*_{\Lambda^*} := \{\, y^* \in Y^* \mid \Lambda^* y^* \in \mathcal{V} \,\},$$

then (1.1) can be rewritten in the form

$$(y^*, \Lambda w) = (\Lambda^* y^*, w)_{\mathcal{V}}, \qquad \forall w \in V. \tag{1.2}$$

We consider the following class of variational problems: find $u \in V$ such that

$$J(u) = \inf \mathcal{P} := \inf_{v \in V} J(v) \qquad (Problem\, \mathcal{P}), \tag{1.3}$$

where

$$J(v) = G(\Lambda v) + F(v), \tag{1.4}$$

the functionals $G : Y \to \mathbb{R}$ and $F : V \to \mathbb{R}$ are convex and lower semicontinuous functionals such that $J(v)$ is a proper functional (cf. [2]) and

$$J(v) \to +\infty \quad \text{as } \|v\|_V \to +\infty. \tag{1.5}$$

In addition, we assume that F is finite at zero element of V and G is coercive on Y.

As usual, the functionals dual to F and G are defined by the relations

$$F^*(v^*) = \sup_{v\in V} \{ < v^*, v > -F(v)\}$$

and

$$G^*(y^*) = \sup_{y\in Y} \{ \, (y^*, y) \, - \, G(y) \, \} \, ,$$

respectively.

If $v^* \in H^*_{\Lambda^*}$, then the first relation admits another form

$$F^*(v^*) = \sup_{v\in V} \{(v^*, v)_{\mathcal{V}} - F(v)\} \, .$$

Existence of a minimizer u to Problem $\mathcal{P}$ follows from standard arguments of the variational calculus (see, e.g., [1, 2]).

Problem $\mathcal{P}$ has a saddle point formulation associated with the Lagrangian

$$L(v, y^*) := \, F(v) + (\, y^*, \Lambda v) \, - G^*(y^*),$$

which is convex and lower semicontinuous with respect to the variable v and concave and upper semicontinuous with respect to the variable y^*.

The Lagrangian yields a dual variational functional defined by the relation

$$\begin{aligned} I^*(y^*) = \inf_{v\in V} L(v, y^*) &= - \, G^*(y^*) + \inf_{v\in V}((y^*, \Lambda v) + F(v)) \\ &= - \, G^*(y^*) - \sup_{v\in V}(< -\Lambda^* y^*, v > \, -F(v)) \\ &= - \, G^*(y^*) - F^*(-\Lambda^* y^*) \end{aligned}$$

and a new (dual) variational problem: find $p^* \in Y^*$ such that

$$I^*(p^*) = \sup_{y^*\in Y^*} \{- \, G^*(y^*) - F^*(-\Lambda^* y^*)\} \qquad (Problem \, \mathcal{P}^*).$$

It is not difficult to show that under the above made assumptions

$$\inf \mathcal{P} \, = \, \sup \mathcal{P}^* := \sup_{y^*\in Y^*} \inf_{v\in V} L(v, y^*) \tag{1.6}$$

and Problem $\mathcal{P}^*$ also has a solution.

2 General form of error identities for convex variational problems

Since both primal and dual problems are well posed and have solutions u^* and p^*, respectively, the pair (u, p^*) is a saddle point of L on $V \times Y^*$, i.e.

$$L(u, y^*) \le L(u, p^*) \le L(v, p^*), \quad \forall v \in V,\ y^* \in Y^*, \tag{2.1}$$

The left–hand side of the inequality yields the relation

$$(y^* - p^*, \Lambda u) \le G^*(y^*) - G^*(p^*), \quad \forall y^* \in Y^*,$$

which means that

$$\Lambda u \in \partial G^*(p^*) \quad \Leftrightarrow \quad p^* \in \partial G(\Lambda u). \tag{2.2}$$

Analogously, the right–hand side of (2.1) yields the relation

$$F(v) - F(u) \ge (p^*, \Lambda(u - v)) = < -\Lambda^* p^*, v - u >, \tag{2.3}$$

which means that

$$-\Lambda^* p^* \in \partial F(u) \quad \Leftrightarrow \quad u \in \partial F^*(-\Lambda^* p^*). \tag{2.4}$$

In general, the relations (2.2) and (2.4) present necessary conditions for the solution pair (u, p^*) and have the form of differential inclusions. However, there is another equivalent way to present these conditions, which is more convenient for our purposes. It is well known (see, e.g., [2, 5]) that (2.2) and (2.4) are equivalent to the relations

$$D_G(\Lambda u, p^*) := G(\Lambda u) + G^*(p^*) - (p^*, \Lambda u) = 0, \tag{2.5}$$

and

$$D_F(u, -\Lambda^* p^*) := F(u) + F^*(-\Lambda^* p^*) + < \Lambda^* p^*, u >= 0, \tag{2.6}$$

respectively.

The functionals $D_G(y^*, y) : Y^* \times Y \to \mathbb{R}$ and $D_F(v^*, v) : V^* \times V \to \mathbb{R}$ (in the literature, they are often called compound functionals) vanish if and only if the arguments satisfy (2.2) and (2.4). In all other cases, they are positive.

Let $q^* \in Y^*$ and $v \in V$ be the functions compared with p^* and u. We introduce the following nonlinear measure of the distance between $\{u, p^*\}$ and $\{v, y^*\}$:

$$\begin{aligned}\mathbb{M}(\{u,p^*\},\{v,y^*\}) :=& \\ = D_F(u,-\Lambda^* y^*) + D_G(\Lambda u, y^*) + D_F(v,-\Lambda^* p^*) + D_G(\Lambda v, p^*).& \quad (2.7)\end{aligned}$$

It is easy to see that $\mathbb{M}(\{u, p^*\}, \{v, y^*\})$ is nonnegative and vanishes if and only if

$$\begin{aligned}\Lambda v \in \partial G^*(p^*), \qquad & y^* \in \partial G(\Lambda u),\\ -\Lambda^* y^* \in \partial F(u), \qquad & v \in \partial F^*(-\Lambda^* p^*).\end{aligned}$$

In other words, $\mathbb{M}(\{u, p^*\}, \{v, y^*\})$ vanishes if and only if all the necessary saddle point conditions are satisfied. Moreover, it was proved (see [5], Section 7.2 and [10]) that

$$\mathbb{M}(\{u,p^*\},\{v,y^*\}) = \; J(v) - I^*(y^*). \qquad (2.8)$$

We see that $\mathbb{M}\{(u, p^*), (v, y^*)\} = 0$ if and only if $J(v) = I^*(y^*)$ what is possible only if v is a minimizer of the problem $\mathcal{P}$ and y^* is a maximizer of the problem $\mathcal{P}^*$. In view of this fact, in [11] the functional $\mathbb{M}$ was introduced as the right error measure for the class of variational problems (1.3)–(1.4). Since any numerical procedure is focused (explicitly or implicitly) on minimization of the duality gap $J(v) - I^*(y^*)$, it automatically minimizes the distance between $\{u, p^*\}$ and $\{v, y^*\}$ in terms of the measure $\mathbb{M}$.

Now we can formulate the main result, which presents the general *a posteriori error identity* for the considered class of problems.

Theorem 2.1 *Let u be a minimizer of the Problem $\mathcal{P}$ and p^* be a maximizer of the Problem $\mathcal{P}^*$. Then, for any $v \in V$ and $y^* \in Y^*$ the following identity holds:*

$$\mathbb{M}\{(u,p^*),(v,y^*)\} \;=\; D_F(v,-\Lambda^* y^*) + D_G(\Lambda v, y^*). \qquad (2.9)$$

The statement directly follows from (2.8). Indeed,

$$\begin{aligned}J(v) - I^*(y^*) &= G(\Lambda v) + F(v) + G^*(y^*) + F^*(-\Lambda^* y^*)\\ &= D_G(\Lambda v, y^*) + (y^*, \Lambda v) + D_F(v,-\Lambda^* y^*) - < \Lambda^* y^*, v > .\end{aligned}$$

We apply (1.1) and arrive at (2.9).

We note that a somewhat different notation the identity (2.9) was proved in [5] (see 7.2.14). It has a clear meaning: the distance between the pair of exact solutions and their approximations measured in terms of the measure $\mathbb{M}$ is equal to the sum of two fully computable functionals $D_G(\Lambda v, y^*)$ and $D_F(v, -\Lambda^* y^*)$ that depend only on approximate solutions and does not contain unknown exact solutions. Therefore, this relation can be viewed as the basic *a posteriori error identity* .

Remark 2.1 It is commonly accepted that errors should be measured in terms of relative (normalized) quantities, which adjust absolute values of errors to a certain measure (e.g., norm) of the exact solution. The relation (2.8) clearly suggests a proper normalization. Since the duality gap $J(v) - I^*(y^*)$ is equal to the error measure $\mathbb{M}(\{u, p^*\}, \{v, y^*\})$ and $C^* := |J(u)| = |I^*(p^*)|$ is a number inside it related to the exact values of the primal/dual energy functionals, it is natural to use the quantity

$$\mathcal{E}(v, y^*) = \frac{1}{C^*}\mathbb{M}(\{u, p^*\}, \{v, y^*\})$$

as a normalized measure of the error (trivial solutions with zero energy are excluded from this consideration). Since $J(u)$ is generally unknown, in practice it may be suggested to use the constant $\widetilde{C}^* = \frac{1}{2}(|J(v)| + |I^*(y^*)|)$ instead of C^*. Then we recall (2.9) and introduce the quantity

$$\widetilde{\mathcal{E}}(v, y^*) = \frac{1}{\widetilde{C}^*}\left(D_F(v, -\Lambda^* y^*) + D_G(\Lambda v, y^*)\right)$$

as a fully computable normalized measure that objectively quantify the accuracy of (v, y^*).

A special, but important case

$$F(v) = < \ell^*, v >, \qquad \ell^* \in V^*$$

deserves a special consideration. We have

$$F^*(-\Lambda^* y^*) = \sup_{v \in V} < -\Lambda^* y^* - \ell^*, v > = \begin{cases} 0 & \text{if } y^* \in Q^*_{\ell^*}, \\ +\infty & \text{if } y^* \notin Q^*_{\ell^*}, \end{cases}$$

where

$$Q^*_{\ell^*} := \{q^* \in Y^* \mid (q^*, \Lambda w) + \langle \ell^*, w \rangle = 0, \quad w \in V\}.$$

Hence,

$$I^*(y^*) = \begin{cases} -G^*(y^*) & \text{if } y^* \in Q^*_{\ell *}, \\ -\infty & \text{if } y^* \notin Q^*_{\ell *} \end{cases}$$

Problem $\mathcal{P}^*$ has the form: find $p^* \in Q^*_{\ell *}$ such that the functional $-G^*(p^*)$ attains its supremum on $Q^*_{\ell *}$.

It is easy to see that the identity (2.8) holds in the form $+\infty = +\infty$ if $y^* \notin Q^*_{\ell *}$ and in the form

$$D_G(\Lambda u, y^*) + D_G(\Lambda v, p^*) = J(v) - I^*(y^*) \tag{2.10}$$

for $y^* \in Q^*_{\ell *}$. Therefore, we conclude that for $v \in V$ and $y^* \in Q^*_{\ell *}$ the error measure is defined by the relation

$$\mathbb{M}\{(u, p^*), (v, y^*)\} = D_G(\Lambda u, y^*) + D_G(\Lambda v, p^*).$$

and the a posteriori error identity (2.9) holds on the affine manifold $Q^*_{\ell *}$ in the form

$$\mathbb{M}\{(u, p^*), (v, y^*)\} = D_G(\Lambda v, y^*). \tag{2.11}$$

Identities (2.10) and (2.11) have been established in [9, 10] and used for the derivation of functional type a posteriori error estimates for a wide class of convex variational problems.

Now we consider particular forms of (2.8)–(2.11) related to some classes of functionals commonly used in mathematical modeling.

3 Problems with quadratic $G(y^*)$

Let U be a Hilbert space endowed with the scalar product $(\cdot, \cdot)_U$ containing the same elements as Y and Y^* and $A : U \to U$ be a bounded, linear, and positive definite operator. The spaces Y and Y^* are identified by the norms

$$\|\!|\tau|\!\|^2 = (A\tau, \tau)_U \quad \text{and} \quad \|\!|\tau|\!\|^2_* = (A^{-1}\tau, \tau)_U,$$

respectively (it is clear that these norms are equivalent to the original norm of U). We define Λ as a linear bounded operator acting from V to U. The conjugate operator $\Lambda^* : U \to V^*$ is defined by the relation

$$(y, \Lambda v)_U = < \Lambda^* y, v > .$$

Consider first the problem

$$\Lambda^* A \Lambda u + \alpha u = \ell^*, \tag{3.1}$$

where α is a positive constant and $\ell^* \in \mathcal{V}$. The corresponding generalized solution u is defined by the relation

$$(A\Lambda u, \Lambda w)_U + \alpha (u, w)_{\mathcal{V}} \;=\; (\ell^*, w)_{\mathcal{V}} \qquad \forall w \in V. \tag{3.2}$$

In this case,

$$G(y) = \frac{1}{2}(Ay, y)_U, \quad G^*(y^*, y^*) = \frac{1}{2}(A^{-1}y^*, y^*)_U,$$

and

$$F(v) = \frac{\alpha}{2}\|v\|^2_{\mathcal{V}} - (\ell^*, v)_{\mathcal{V}}.$$

We find that for any $v^* \in \mathcal{V}$

$$F^*(v^*) = \sup_{v \in V} \Big\{ (v^* + \ell^*, v)_{\mathcal{V}} - \frac{\alpha}{2}\|v\|^2_{\mathcal{V}} \Big\} = \frac{1}{2\alpha}\|v^* + \ell^*\|^2_{\mathcal{V}}.$$

For any $y^* \in Y^*$, we have

$$\begin{aligned} D_G(\Lambda u, y^*) = \frac{1}{2}\left((A\Lambda u, \Lambda u)_U + (A^{-1}y^*, y^*)_U - 2(\Lambda u, y^*)_U \right) \\ = \frac{1}{2}\|\!| A\Lambda u - y^* |\!\|^2_* \end{aligned} \tag{3.3}$$

and

$$D_G(\Lambda v, p^*) = \frac{1}{2}\|\!| A\Lambda v - p^* |\!\|^2_*. \tag{3.4}$$

Let $y^* \in H^*_{\Lambda^*}$. Then,

$$\begin{aligned} D_F(u, -\Lambda^* y^*) = \frac{\alpha}{2}\|u\|^2_{\mathcal{V}} + \frac{1}{2\alpha}\|\ell^* - \Lambda^* y^*\|^2_{\mathcal{V}} + (u, \Lambda^* y^* - \ell^*)_{\mathcal{V}} \\ = \frac{1}{2\alpha}\|\Lambda^* y^* + \alpha u - \ell^*\|^2_{\mathcal{V}} \end{aligned} \tag{3.5}$$

and quite analogously (note that $p^* \in H^*_{\Lambda^*}$) we obtain

$$D_F(v, -\Lambda^* p^*) = \frac{1}{2\alpha}\|\Lambda^* p^* + \alpha v - \ell^*\|^2_{\mathcal{V}}. \tag{3.6}$$

Now we recall (2.2) and (2.4). Since the functionals G and G^* are Gateaux differentiable, the relations (2.2) have the form

$$p^* = G'(\Lambda u) = A\Lambda u \quad \text{and} \quad \Lambda u = (G^*)'(p^*) = A^{-1}p^*. \tag{3.7}$$

The functionals F, and F^* are also differentiable. Therefore, (2.4) have the form

$$u = (F^*)'(-\Lambda^* p^*) = \frac{1}{\alpha}(\ell^* - \Lambda^* p^*) \tag{3.8}$$

and

$$-\Lambda^* p^* = F'(u) = \alpha u - \ell^*. \tag{3.9}$$

By (3.7)–(3.9) we conclude that the components of the measure $\mathbb{M}$ are as follows:

$$D_G(\Lambda u, y^*) = \quad \frac{1}{2}|\!|\!| p^* - y^* |\!|\!|_*^2, \tag{3.10}$$

$$D_G(\Lambda v, p^*) = \quad \frac{1}{2}|\!|\!| \Lambda(u - v) |\!|\!|^2, \tag{3.11}$$

$$D_F(u, -\Lambda^* y^*) = \quad \frac{1}{2\alpha}\|\Lambda^*(y^* - p^*)\|_{\mathcal{V}}^2, \tag{3.12}$$

$$D_F(v, -\Lambda^* p^*) = \quad \frac{\alpha}{2}\|v - u\|_{\mathcal{V}}^2. \tag{3.13}$$

Thus, for this class of linear problems the measure $\mathbb{M}\{(u, p^*), (v, y^*)\}$ is defined by the sum of above presented four norms of two error functions $e := u - v$ and $\eta^* := p^* - y^*$.

It is easy to see that $\mathbb{M}\{(u, p^*), (v, y^*)\}$ is equivalent to the sum of two norms associated with the primal and dual errors:

$$|\!|\!| e |\!|\!|_\alpha^2 := \frac{1}{2}\left(|\!|\!| \Lambda e |\!|\!|^2 + \alpha\|e\|_{\mathcal{V}}^2\right). \tag{3.14}$$

and

$$\|\eta^*\|_{H^*, \frac{1}{\alpha}}^2 := \frac{1}{2}\left(|\!|\!| \eta^* |\!|\!|_*^2 + \frac{1}{\alpha}\|\Lambda^* \eta^*\|_{\mathcal{V}}^2\right) \tag{3.15}$$

Here the first norm is the energy norm associated with the primal variational functional J and the second one can be viewed a norm of the space $H^*_{\Lambda^*}$. We see that

$$\mathbb{M}\left(\{u, p^*\}, \{v, y^*\}\right) \; = \; |\!|\!| e |\!|\!|_\alpha^2 + \|\eta^*\|_{H^*, \frac{1}{\alpha}}^2$$

and the identity (2.8) reads

$$\|\!|e|\!\|_\alpha^2 + \|\eta^*\|^2_{H^*,\frac{1}{\alpha}} = J(v) - I^*(y^*). \tag{3.16}$$

In other words, for this class of variational problems the difference between the primal and dual functionals measured in terms of $\mathbb{M}$ is equal to the sum of specially selected norms.

Theorem 2.1 implies the following a posteriori error identity:

$$\|\!|e|\!\|_\alpha^2 \,+\, \|\eta^*\|^2_{H^*,\frac{1}{\alpha}} \;=\; \frac{1}{2}\|\!|A\Lambda v \,-\, y^*|\!\|_*^2 \,+\, \frac{1}{2\alpha}\|\!|\Lambda^* y^* \,+\, \alpha v \,-\, \ell^*|\!\|_{\mathcal{V}}^2. \tag{3.17}$$

The right hand side of this identity contains only known functions and vanishes if and only if

$$\begin{aligned} &A\Lambda v - y^* = 0,\\ &\Lambda^* y^* + \alpha v - \ell^* = 0, \end{aligned}$$

i.e., if $v = u$ (cf. (3.1)) and $y^* = p^*$. In all other cases the right hand side is positive and equals to the combined primal–dual measure of the error presented by the left hand side. We note that such type identity (both sides of which are expressed in terms of squared norms) takes place only for this class of linear problems.

The identity (3.17) is not valid for $\alpha = 0$. In this case, then we must use (2.10) and (2.11) and use introduce the condition

$$\Lambda^* y^* = \ell^*. \tag{3.18}$$

We find that

$$\begin{aligned} \mathbb{M}\{(u,p^*),(v,y^*)\} \;&= \frac{1}{2}\|\!|A\Lambda u - y^*|\!\|_*^2 + \frac{1}{2}\|\!|A\Lambda v - p^*|\!\|_*^2 \\ &= \frac{1}{2}\|\!|\eta^*|\!\|^2 + \frac{1}{2}\|\!|\Lambda e\,|\!\|^2. \end{aligned}$$

Then, (2.10) yields the identity

$$\frac{1}{2}\|\!|\eta^*|\!\|_*^2 + \frac{1}{2}\|\!|\Lambda e\,|\!\|^2 = J(v) - I^*(y^*). \tag{3.19}$$

Set here $y^* = p^*$. Since

$$I^*(p^*) = J(u) \quad \text{and} \quad p^* = A\Lambda u,$$

we obtain

$$\frac{1}{2}\|\Lambda e\|^2 \;=\; J(v) - J(u). \tag{3.20}$$

This is a generalized form of the Mikhlin's error identity (see, e.g., [4, 3]), which was derived for variational functionals defined by quadratic forms.

By (2.11), we find that

$$\|A\Lambda u - y^*\|_*^2 + \|A\Lambda v - p^*\|_*^2 = \|A\Lambda v - y^*\|_*^2. \tag{3.21}$$

We can rewrite it in an equivalent form

$$\|\eta^*\|_*^2 \,+\, \|\Lambda e\|^2 \;=\; \|A\Lambda v - y^*\|_*^2, \quad \forall y^* \in Q^*_{\ell^*}. \tag{3.22}$$

The latter identity can be viewed as a generalization of the Prager–Synge error relation derived in [7] for linear elasticity problems.

Remark 3.1 We see that error measures arising in the estimates are generated by nonnegative compound functionals. In the linear case (i.e., for quadratic type functionals), they are equivalent to norms. However in general, $\mathbb{M}$ consists of nonlinear terms that jointly form a proper measure of the accuracy (see [11]).

4 Particular cases

Now we briefly discuss applications of the above presented error identities to particular classes variational problems.

4.1 Quadratic growth problems with $\Lambda = \operatorname{grad}$

Let $V = H_0^1(\Omega)$, where Ω is a bounded Lipschitz domain in $\mathbb{R}^d$ ($d \geq 1$). We set $U = L^2(\Omega, \mathbb{R}^d)$, $\mathcal{V} = L^2(\Omega)$, and identify A with a symmetric real matrix in $\mathbb{M}^{d\times d}_{sym}$. Then $\Lambda^* = -\operatorname{div}$ and (3.1) is the equation

$$\operatorname{div} A\nabla u - \alpha u + \ell^* = 0. \tag{4.1}$$

In this case,

$$\|e\|_\alpha^2 = \frac{1}{2}\int\limits_\Omega \left(A\nabla e\cdot\nabla e + \alpha|e|^2\right) dx,$$

$$\|\eta^*\|^2_{H^*,\frac{1}{\alpha}} = \frac{1}{2}\int\limits_\Omega \left(A^{-1}\eta^*\cdot\eta^* + \frac{1}{\alpha}|\operatorname{div}\eta^*|^2\right),$$

and the a posteriori error identity (3.17) has the form

$$|||e|||_{\alpha}^2 + |||\eta^*|||_{H^*,\frac{1}{\alpha}}^2 = \frac{1}{2}\int\limits_{\Omega} \left(A\nabla v\cdot\nabla v + A^{-1}y^*\cdot y^* - y^*\cdot\nabla v\right) dx + \frac{1}{2\alpha}\|\operatorname{div} y^* - \alpha v + \ell^*\|^2. \quad (4.2)$$

If $\alpha = 0$, then we use (3.18) and (3.19) and obtain the identities

$$\frac{1}{2}\int\limits_{\Omega} A^{-1}\eta^*\cdot\eta^*\,dx + \frac{1}{2}\int\limits_{\Omega} A\nabla e\cdot\nabla e\,dx = J(v) - I^*(y^*) \quad (4.3)$$

and the Mikhlin's identity

$$\frac{1}{2}\int\limits_{\Omega} A\nabla e\cdot\nabla e\,dx \;=\; J(v) - J(u). \quad (4.4)$$

By (3.22) we obtain a version of the Prager-Synge identity

$$\int\limits_{\Omega}(A^{-1}\eta^*\cdot\eta^* + A\nabla e\cdot\nabla e)dx = \int\limits_{\Omega}\left(A\nabla v\cdot\nabla v + A^{-1}y^*\cdot y^* - \nabla v\cdot y^*\right)dx. \quad (4.5)$$

4.2 Problems with the operator $\Lambda = \operatorname{Sym}\nabla$

Problems of this type arise in continuum media problems, where Λ is a symmetric part of the tensor ∇u and u is a vector field. In this case, the error identities are quite similar to (4.2)–(4.5). The reader can find a sistematic discussion of them and respective error majorants in [5, 9, 10, 12]).

4.3 Generalized Stokes problem

If V coincides with the space $S_0^{1,2}(\Omega,\mathbb{R}^d)$ that is the closure of smooth solenoidal fields with respect to the norm of $H^1(\Omega,\mathbb{R}^d)$ and $\Lambda v = \nabla v$, where v is the velocity vector field, then we arrive at a class of variational problems generated by incompressible media. The generalized Stokes problem is one of the most known problems in this class. It often arises in time

discretization of the parabolic Stokes problem. It is related to the system

$$-\nu\Delta u + \alpha u = \ell^* \qquad \text{in } \Omega, \tag{4.6}$$

$$\operatorname{div} u = 0, \tag{4.7}$$

$$u = u_0 \qquad \text{on } \partial\Omega, \tag{4.8}$$

where u_0 is a divergence free field in $H^1(\Omega, \mathbb{R}^d)$ and $\nu > 0$ is the viscosity.

In this case, $\Lambda^* = -\mathrm{Div}$ (i.e., the conjugate operator is formed by the divergence of a tensor field), $U = L^2(\Omega, \mathbb{R}^d)$,

$$G(\Lambda v) = \int\limits_\Omega \frac{\nu}{2}|\nabla v|^2 dx, \quad \text{and} \quad G^*(y^*) = \int\limits_\Omega \frac{1}{2\nu}|y^*|^2 dx.$$

Here $|y^*|$ denotes the Euclidean norm of the tensor y^* ($|y^*|^2 := y^* : y^*$).

Let $v \in S_0^{1,2}(\Omega, \mathbb{R}^d)$ and $y^* \in L^2(\Omega, \mathbb{M}^{d\times d}_{sym})$ be approximations of the exact velocity u and exact stress σ^*, respectively. Then the general method exposed in Sect. 2 suggests to measure the errors $e = u - v$ and $\eta^* = \sigma^* - y^*$ (for the velocity and stress) in terms of the integral type measures

$$\|\!|e|\!\|^2_\alpha = \frac{1}{2}\int\limits_\Omega \left(\nu|\nabla e|^2 + \alpha|e|^2\right) dx$$

and

$$\|\!|\eta^*|\!\|^2_{H^*,\frac{1}{\nu}} = \frac{1}{2}\int\limits_\Omega \left(\frac{1}{\nu}|\eta^*|^2 + \frac{1}{\alpha}|\operatorname{Div}\eta^*|^2\right) dx,$$

respectively.

We conclude that the a posteriori error identity (3.17) has the form

$$\|\!|e|\!\|^2_\alpha + \|\!|\eta^*|\!\|^2_{H^*,\frac{1}{\nu}} = \frac{1}{2}\int\limits_\Omega \left(\nu|\nabla v|^2 + \frac{1}{\nu}|y^*|^2 - y^* : \nabla v\right) dx + \frac{1}{2\alpha}\|\operatorname{Div} y^* - \alpha v + \ell^*\|^2. \tag{4.9}$$

4.4 Nonlinear problem

Finally, we consider an example of highly nonlinear problem, where G is a power growth functional and F has linear growth with respect to v. Let

$$J(v) = \frac{1}{q}\int\limits_\Omega |\nabla v|^q\, dx + \alpha\int\limits_\Omega |v|\, dx + \int\limits_\Omega fv dx.$$

We assume that $0 < q < +\infty$, $\alpha > 0$, and f is a bounded real valued function. Existence of the minimizer u is obvious because $J(v)$ is coercive of the reflexive space $V = \overset{\circ}{W}{}^{1,q}(\Omega)$. In this case, $\Lambda = \nabla$, $\Lambda^* = -\mathrm{div}\,$,

$$G^*(y^*) = \frac{1}{q^*} \int\limits_\Omega |y^*|^{q^*}\, dx, \quad \frac{1}{q} + \frac{1}{q^*} = 1,$$

and

$$D_G(y, y^*) = \int\limits_\Omega \left(\frac{1}{q}|y|^q + \frac{1}{q^*}|y^*|^{q^*} - yy^* \right) dx.$$

Next,

$$F(v) = \alpha \int\limits_\Omega |v| dx + \int\limits_\Omega fv dx,$$

For any real valued function $v^* \in V^*$, we have

$$F^*(v^*) = \sup_{v \in V} \int\limits_\Omega ((v^* - f)v - \alpha|v|) dx$$

$$= \begin{cases} 0 & \text{if } |v^* - f| \le \alpha, \\ +\infty & \text{if } |v^* - f| > \alpha \end{cases} \qquad (4.10)$$

and, therefore,

$$F^*(-\Lambda^* y^*) = \begin{cases} 0 & \text{if } |\mathrm{div}\, y^* + f| \le \alpha, \\ +\infty & \text{if } |\mathrm{div}\, y^* + f| > \alpha. \end{cases}$$

Hence

$$D_F(v, -\Lambda^* y^*) = \begin{cases} \displaystyle\int\limits_\Omega (\alpha|v| + v(\mathrm{div}\, y^* + f)) dx & \text{if } |\mathrm{div}\, y^* + f| \le \alpha, \\ +\infty & \text{if } |\mathrm{div}\, y^* + f| > \alpha. \end{cases}$$

We see that the measure $\mathbb{M}$ is finite only if

$$y^* \in Q^*_\alpha := \{ y^* \in Y^* \mid |\mathrm{div}\, y^*(x) + f| \le \alpha \quad \text{for a.a. } x \in \Omega \}. \qquad (4.11)$$

This condition plays the same role as (3.18) for variational problems with $F(v) =< \ell^*, v >$. However, there is an essential difference. Now the error

identities are finite not on the set $Q^*_{\ell *}$ (which is an affine manifold defined by (3.18)) but in a "strip" Q^*_α and width of this strip depends on the parameter α.

Note that $p^* = |\nabla u|^{q-2}\nabla u$ and $\nabla u = |p^*|^{\alpha^*-2}p^*$. Another duality relation reads

$$\operatorname{div} p^* + f = \begin{cases} -\alpha \frac{u}{|u|} & \text{if } u \neq 0, \\ -\alpha\zeta \quad \text{where } |\zeta| \leq 1 & \text{if } u = 0 \end{cases}$$

and we conclude that $p^* \in Q^*_\alpha$. In view of (2.7), for any $y^* \in Q^*_\alpha$ and $v \in V$, the measure $\mathbb{M}$ is defined by the relation

$$\begin{aligned} &\mathbb{M}(\{u, p^*\}, \{v, y^*\}) \\ &\qquad = \int_\Omega (\alpha|u| + u(\operatorname{div} y^* + f) + \alpha|v| + v(\operatorname{div} p^* + f))\, dx \\ &\qquad\quad + \int_\Omega \left(\frac{1}{q}|p^*|^q + \frac{1}{q^*}|y^*|^{q^*} - p^* \cdot y^* |p^*|^{\alpha^*-2}\right) dx \\ &\qquad\quad + \int_\Omega \left(\frac{1}{q}|\nabla v|^q + \frac{1}{q^*}|\nabla u|^q - \nabla v \cdot \nabla u |\nabla u|^{q-2}\right) dx. \end{aligned} \tag{4.12}$$

It is easy to see that the measure vanishes if $u = v$ and $p^* = y^*$. Now Theorem 2.1 yields the following error identity for this variational problem:

$$\begin{aligned} \mathbb{M}(\{u, p^*\}, \{v, y^*\}) &= \int_\Omega (\alpha|v| + v(\operatorname{div} y^* + f))\, dx \\ &\quad + \int_\Omega \left(\frac{1}{q}|\nabla v|^q + \frac{1}{q^*}|y^*|^{q^*} - \nabla v \cdot y^*\right) dx. \end{aligned} \tag{4.13}$$

Finally, we note that the problem considered above generates an elliptic variational inequality of the second kind. Analysis of suitable error measures (and corresponding error majorants) for variational inequalities of the first kind is presented in [13] for obstacle type problems and in [6] for problems with nonlinear boundary conditions.

References

[1] V. Barbu and T. Precupanu. Convexity and optimization in Banach spaces. Fourth edition. Springer Monographs in Mathematics. Springer, Dordrecht, 2012.

[2] I. Ekeland and R. Temam. *Convex analysis and variational problems* North-Holland, Amsterdam, 1976.

[3] H. Gajewskii, K. Gröger, and K. Zacharias, *Nichtlineare Operatorgliechungen and Operatordifferentialgleiichungen,* Akademie-Verlag, Berlin, 1974.

[4] S. G. Mikhlin. Variational methods in mathematical physics. Pergamon, Oxford, 1964.

[5] P. Neittaanmäki and S. Repin. Reliable methods for computer simulation. Error control and a posteriori estimates. Studies in Mathematics and its Applications, 33. Elsevier Science B.V., Amsterdam, 2004.

[6] P. Neittaanmäki, S. Repin, and J. Valdman. Estimates of deviations from exact solutions of elasticity problems with nonlinear boundary conditions. *Russian J. Numer. Anal. Math. Modeling*, 28(2013), 6, 597–630.

[7] W. Prager and J. L. Synge, *Approximation in elasticity based on the concept of function space,* Quart. Appl. Math. 5(1947), pp. 241-269.

[8] S. Repin, *A posteriori error estimation for nonlinear variational problems by duality theory,* Zapiski Nauchn, Semin, V.A. Steklov Mathematical Institute in St.-Petersburg (POMI), 243(1997), pp. 201-214.

[9] S. Repin, *A posteriori error estimates for variational problems with uniformly convex functionals,* Math. Comput., 69(230), 2000, pp. 481-500.

[10] S. Repin, *Two-sided estimates of deviation from exact solutions of uniformly elliptic equations,* Proc. St. Petersburg Math. Society, IX(2001), pp. 143–171, translation in Amer. Math. Soc. Transl. Ser. 2, 209, Amer. Math. Soc., Providence, RI, 2003.

[11] S. Repin. On measures of errors for nonlinear variational problems. Russian J. Numer. Anal. Math. Modelling 27 (2012), no. 6, 577?–584.

[12] S. Repin, A posteriori estimates for partial differential equations, Walter de Gruyter, Berlin, 2008.

[13] S. Repin and J. Valdman. A posteriori error estimates for two–phase obstacle problem, J. Math. Sci., 207 (2015), 2, 324–335.

Ann. Acad. Rom. Sci.
Ser. Math. Appl.
ISSN 2066-6594 Vol. 7, No. 1/2015

POD-DEIM APPROACH ON DIMENSION REDUCTION OF A MULTI-SPECIES HOST-PARASITOID SYSTEM*

Gabriel Dimitriu† Răzvan Ştefănescu‡ Ionel M. Navon§

Dedicated to the memory of Prof. Dr. Viorel Arnăutu

Abstract

In this study, we implement the DEIM algorithm (Discrete Empirical Interpolation Method) combined with POD (Proper Orthogonal Decomposition) to provide dimension reduction of a model describing the aggregative response of parasitoids to hosts in a coupled multi-species system. The model is defined by five reaction-diffusion-chemotaxis equations. We show DEIM improves the efficiency of the POD approximation and achieves a complexity reduction of the nonlinear terms. Numerical results are presented.

MSC: 35K57, 65M06, 65Y20.

keywords: reduced order modeling, proper orthogonal decomposition, discrete empirical interpolation method.

*Accepted for publication in revised form on January 26-th, 2015

†dimitriu.gabriel@gmail.com, "Grigore T. Popa" University of Medicine and Pharmacy, Department of Mathematics and Informatics, Iaşi 700115, Romania

‡Virginia Polytechnic Institute and State University, Computer Science Department, Blacksburg, VA 24060, USA

§The Florida State University, Department of Scientific Computing, Tallahassee, FL 32306, USA

1 Introduction

Reduced order modeling refers to the development of low-dimensional models that represent the important characteristics of a high-dimensional dynamical system. Typically, reduced models are constructed by projecting the high-fidelity model onto a suitably chosen low-dimensional subspace ([1]). While for linear models it is possible to produce input-independent high accurate reduced models, in the case of general nonlinear systems the transfer function approach is not applicable and input-specified semi-empirical methods are usually employed. Most approaches for nonlinear problems construct the reduced bases from a collection of simulations (method of snapshots [21, 22, 23]).

Proper Orthogonal Decomposition (POD) – see [3, 5, 9, 10, 15, 25] and the references therein – is probably the mostly used and most successful model reduction technique, where the basis functions contain information from the solutions of the dynamical system at pre-specified time-instances, so-called snapshots. Due to a possible linear dependence or almost linear dependence, the snapshots themselves are not appropriate as a basis. Instead two methods can be employed, singular value decomposition (SVD) for the matrix of snapshots or eigenvalue decomposition for the correlation matrix ([24]). The singular value decomposition based POD basis construction is more computationally efficient since it decomposes the snapshots matrix whose condition number is the square root of the correlation matrix used in the eigenvalue decomposition.

Tensorial POD for reducing the computational complexities of the nonlinearity terms was traditionally employed in POD Galerkin by the fluid mechanics community ([23, 15, 16]), and a matrix formulation named *pre-computing technique* was introduced in [6] for calculation of quadratic nonlinearities. An extension of the tensorial based calculus to compute POD Galerkin p^{th} order polynomial nonlinearities has been proposed in [25].

A considerable reduction in the nonlinear terms complexity is achieved by DEIM ([7]) – a discrete variation of Empirical Interpolation Method (EIM), proposed by Barrault, Maday, Nguyen and Patera in [4]. According to this method, the evaluation of the approximate nonlinear term does not require a prolongation of the reduced state variables back to the original high dimensional state approximation required to evaluate the nonlinearity in the POD approximation.

Recently the use of interpolation methods relying on greedy algorithms became attractive for computing the reduced order nonlinear terms derivatives. Based on EIM, the Multi-Component Empirical Interpolation Method

([27]) derives affine approximations for continuous vector valued functions, while matrix DEIM (MDEIM) technique ([28]) relies on DEIM to approximate the Jacobian of a nonlinear function to obtain aposteriori error estimates of DEIM reduced nonlinear dynamical system. Significant progress in the construction of implicit reduced order models is provided by the development of the sparse matrix DEIM method ([26]) that uses samples of the nonzero entries of the full Jacobian matrix and thus can approximate very large matrices, unlike the current MDEIM method which is limited by its large computational memory requirements.

In this work, we perform an application of DEIM combined with POD to obtain dimension reduction of a model describing the interactions of the two hosts and two parasitoids in a one-dimensional domain in the presence of a chemotaxis process. The model was introduced and analyzed by Pearce et al. in [19, 20] with respect to the stability properties of the steady-states. The behaviour of the parasitoids towards plant infochemicals generated during host feeding are defined as a chemotactic response and the plant infochemicals are viewed as chemoattractants. The model considers a single chemoattractant produced in proportion to the total host density. Both parasitoids play the role of biological control agents against the hosts.

The paper is organized as follows. Section 2 describes the equations of parasitoid model under study. Section 3 describes the POD and DEIM methods along with Galerkin projection. Results of illustrative numerical experiments are discussed in Section 4 while conclusions are drawn in Section 5.

2 The multi-species host-parasitoid model

We describe here the parameters and the model equations introduced by Pearce et al. in [20]. The reaction kinetics describing the interactions between hosts and parasitoids are coupled with spatial motility and chemotaxis terms giving rise to a system of reaction-diffusion-chemotaxis equations.

In the absence of parasitism, both host species are modelled by logistic, density-dependent growth, with growth rates r_1 and r_2 and carrying capacities K_1 and K_2, respectively. Parasitism by both parasitoids is modelled by an Ivlev functional response. *C. glomerata* parasitises *P. brassicae* at rate α_1 and *P. rapae* at rate α_2. *C. rubecula* parasitises *P. rapae* at rate α_3. The efficiency of parasitoid discovery of hosts is denoted by a_1, a_2 and a_3. Each parasitised host gives rise to e_1, e_2 and e_3 next-generation parasitoids. The parasitoids are subject to mortality rates d_1 (*C. glomerata*) and d_2 (*C.*

rubecula).

The motility coefficients D_1, D_2, D_3 and D_4 of the four species are constants and determine the rate at which each species disperses randomly throughout the domain. The chemoattractant K is generated proportionally to the total host density $(N + M)$ at the rate r_3 and decays at the rate d_3. The motility coefficient of the chemoattractant, D_5, is a constant and defines the rate at which the chemoattractant diffuses through the domain. The chemotactic response of both species of parasitoid is modelled as a linear response and the strength of the response depends on the chemotaxis coefficients χ_1 and χ_2. The model is defined by the equations ([20]):

$$
\begin{aligned}
\frac{\partial N}{\partial t} &= \overbrace{D_1\nabla^2 N}^{\text{random motility}} + \overbrace{r_1 N\left(1-\frac{N}{K_1}\right)}^{\text{logistic growth}} - \overbrace{\alpha_1 P(1-e^{-a_1 N})}^{\text{mortality due to parasitism}}\,, \\
\frac{\partial M}{\partial t} &= D_2\nabla^2 M + r_2 M\left(1-\frac{M}{K_2}\right) - \alpha_2 P(1-e^{-a_2 M}) \\
&\quad -\alpha_3 Q(1-e^{-a_3 M})\,, \\
\frac{\partial P}{\partial t} &= D_3\nabla^2 P - \chi_1\nabla\cdot(P\nabla k) + e_1\alpha_1 P(1-e^{-a_1 N}) \\
&\quad +e_2\alpha_2 P(1-e^{-a_2 M}) - d_1 P\,, \\
\frac{\partial Q}{\partial t} &= \underbrace{D_4\nabla^2 Q}_{\text{random motility}} - \underbrace{\chi_2\nabla\cdot(Q\nabla k)}_{\text{parasitoid chemotactic response}} \\
&\quad + \underbrace{e_3\alpha_3 Q(1-e^{-a_3 M})}_{\text{growth to the parasitism}} - \underbrace{d_2 Q}_{\text{mortality}}\,, \\
\frac{\partial K}{\partial t} &= D_5\nabla^2 K + \underbrace{r_3(N+M)}_{\text{production}} - d_3 K\,,
\end{aligned}
\tag{2.1}
$$

where N and M are the density of hosts *P. brassicae* and *P. rapae*, respectively, P and Q represent the density of parasitoids *C. glomerata* and *C. rubecula*, and K represents the concentration of the chemoattractant produced during feeding by the hosts. $N = N(x,t)$ denotes local population density (organisms per area) at time t and spatial coordinate x (and likewise for M, P, and Q). $k = k(x,t)$ denotes local chemoattractant concentration at time t and spatial coordinate x.

Here we consider the system (2.2) in a bounded domain Ω with smooth boundary $\partial\Omega$ and homogeneous Dirichlet boundary conditions (which correspond to a hostile external habitat). The initial conditions given by

$N(x,0) = N_0(x)$, $M(x,0) = M_0(x)$ $P(x,0) = P_0(x)$,$Q(x,0) = Q_0(x)$ and $K(x,0) = K_0(x)$ will be specified in Section 4.

Using the non-dimensional variables: $t' = r_1 t$, $x' = \frac{x}{L}$, $N' = \frac{N}{K_1}$, $M' = \frac{M}{K_2}$, $P' = \frac{P}{K_1}$, $Q' = \frac{Q}{K_2}$, $K' = \frac{K}{K_0}$, and dropping primes one obtains the nondimensionalised system:

$$\begin{aligned}
\frac{\partial N}{\partial t} &= D_N \nabla^2 N + N\,(1-N) - s_1 P(1-e^{-\rho_1 N})\,, \\
\frac{\partial M}{\partial t} &= D_M \nabla^2 M + \gamma_1 M\,(1-M) - s_2 P(1-e^{-\rho_2 M}) \\
&\quad -s_3 Q(1-e^{-\rho_3 M})\,, \\
\frac{\partial P}{\partial t} &= D_P \nabla^2 P - \chi_P \nabla\cdot(P\nabla k) + c_1 P(1-e^{-\rho_1 N}) \\
&\quad +c_2 P(1-e^{-\rho_2 M}) - \eta_1 P\,, \\
\frac{\partial Q}{\partial t} &= D_Q \nabla^2 Q - \chi_2 \nabla\cdot(Q\nabla k) + c_3 Q(1-e^{-\rho_3 M}) - \eta_2 Q\,, \\
\frac{\partial K}{\partial t} &= D_K \nabla^2 K + \gamma_2(N+\gamma_3 M) - \eta_3 K
\end{aligned} \tag{2.2}$$

where $D_N = \frac{D_1}{r_1 L^2}$, $D_M = \frac{D_2}{r_1 L^2}$, $D_P = \frac{D_3}{r_1 L^2}$, $D_Q = \frac{D_4}{r_1 L^2}$, $D_K = \frac{D_5}{r_1 L^2}$, $\chi_P = \frac{\chi_1 k_0}{r_1 L^2}$, $\chi_Q = \frac{\chi_2 k_0}{r_1 L^2}$, $\rho_1 = \frac{a_1}{K_1}$, $\rho_2 = \frac{a_2}{K_2}$, $\rho_3 = \frac{a_3}{K_2}$, $\gamma_1 = \frac{r_1}{r_2}$, $\gamma_2 = \frac{r_3}{K_1} r_1$, $\gamma_3 = \frac{K_2}{K_1}$, $s_1 = \frac{\alpha_1}{r_1}$, $s_2 = \frac{\alpha_2 K_1}{\alpha_1 K_2}$, $s_3 = \frac{\alpha_3}{r_1}$, $c_1 = \frac{e_1\alpha_1}{r_1}$, $c_2 = \frac{e_2\alpha_2}{r_1}$, $c_3 = \frac{e_3\alpha_3}{r_1}$, $\eta_1 = \frac{d_1}{r_1}$, $\eta_2 = \frac{d_2}{r_1}$ and $\eta_3 = \frac{d_3}{r_1}$.

3 The POD and POD-DEIM reduced order systems

In this section we briefly present some details for constructing the reduced-order system of the full-order system (2.2) applying Proper Orthogonal Decomposition (POD) and Discrete Empirical Interpolation Method (DEIM).

POD is an efficient method for extracting orthonormal basis elements that contain characteristics of the space of expected solutions which is defined as the span of the snapshots ([9, 10, 14, 15]). In this framework, snapshots are the sampled (numerical) solutions at particular time steps or at particular parameter values. POD gives an optimal set of basis vectors minimizing the mean square error of a reduced basis representation.

Our reduced order modeling description uses a discrete inner product though continuous products may be employed too. Generally, an unsteady

model is usually governed by the following semi–discrete dynamical system

$$\frac{d\mathbf{y}(t)}{dt} = \mathbf{F}(\mathbf{y}, t), \quad \mathbf{y}(0) = \mathbf{y}_0 \in \mathbb{R}^n, \quad n \in I\!N, \tag{3.1}$$

n being the number of space points discretizing the domain. From the temporal-spatial flow $\mathbf{y}(t) \in \mathbb{R}^n$, we select an ensemble of N_t time instances $\mathbf{y}_1, ..., \mathbf{y}_{N_t} \in \mathbb{R}^n$, where $N_t \in I\!N,\ N_t > 0$. If we denote by $\bar{\mathbf{y}} = \frac{1}{N}\sum_{i=1}^{n} \mathbf{y}_i$ the mean field correction, one way to compute the POD basis is to apply an eigenvalue decomposition to the correlation matrix $W = [w_{ij}]_{i,j=1,..,N_t}$, $w_{ij} = \langle \mathbf{y}_i - \bar{\mathbf{y}}, \mathbf{y}_j - \bar{\mathbf{y}} \rangle$, where $\langle \cdot, \cdot \rangle$ is the Euclidean dot product. The corresponding eigenvalues are denoted by $\lambda_i \geq 0,\ i = 1, .., N_t$ and the eigenvectors are stored in a matrix $\Phi = [\phi_{ij}]_{i,j=1,..,Nt},\ \Phi \in \mathbb{R}^{N_t \times N_t}$. Then the orthonormal POD basis vectors are computed using $\mathbf{v}_i = \sum_{j=1}^{N_t} \phi_{ij}(\mathbf{y}_i - \bar{\mathbf{y}}),\ i = 1, .., N_t$.

Next, we introduce a relative information content to select a low-dimensional basis of size $k \ll n$, by neglecting modes corresponding to the small eigenvalues. Define $I(m) = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$ and k is chosen such that $k = \min\{I(m) : I(m) \geq \gamma\}$ where $0 \leq \gamma \leq 1$ is larger than 99% of the total kinetic energy captured by the reduced space $V = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k\}$. The way the POD basis is constructed ensures that the mean square error between $\mathbf{y}(t_i)$ and POD expansion $\mathbf{y}^{POD}(t_i) = \bar{\mathbf{y}} + V\tilde{\mathbf{y}}(t_i),\ \tilde{\mathbf{y}}(t_i) \in \mathbb{R}^k$, for all $i = 1, .., N_t$ and $k = 1, .., N_t$ is minimized on average [14, p. 4].

By employing a Galerkin projection, the full model equations (3.1) is projected onto the space V spanned by the POD basis elements and the POD reduced order model is obtained

$$\frac{d\tilde{\mathbf{y}}(t)}{dt} = V^T \mathbf{F}\left(\bar{\mathbf{y}} + V\tilde{\mathbf{y}}(t), t\right), \quad \tilde{\mathbf{y}}(0) = V^T\left(\mathbf{y}(0) - \bar{\mathbf{y}}\right). \tag{3.2}$$

The efficiency of the POD-Galerkin technique is limited to linear or bilinear terms, since the projected nonlinear terms still depend on all the variables of the full model. To mitigate this inefficiency the *discrete empirical interpolation method* (DEIM) [6, 7, 8, 17] and the *empirical interpolation method* (EIM) [4, 13, 18] approximate the nonlinear terms via effective affine offline-online computational decompositions.

The projected nonlinearity in the system (3.2) is approximated by DEIM in the form that enables precomputation, so that evaluating the approximate nonlinear terms using DEIM does not require a prolongation of the reduced state variables back to the original high dimensional state approximation, as it is required for nonlinearity evaluation in the original POD approximation. Only a few entries of the original nonlinear term, corresponding to the specially selected interpolation indices from DEIM must be evaluated at each

time step ([4, 6, 7, 11, 24]). We provide formally the DEIM approximation in Definition 1, and the procedure for selecting DEIM indices is shown in Algorithm DEIM. Each DEIM index is selected to limit the growth of a global error bound for nonlinear terms using a greedy technique ([7]).

Definition 1 *Let* $\{\mathbf{u}_\ell\}_{\ell=1}^m \subset \mathbb{R}^n$ *be a linearly independent set denoted by* $\mathbf{U}$*, which is computed from the snapshots of the nonlinear term* $\mathbf{F}$ *in (3.1). The* DEIM *approximation of order* m *for* $\mathbf{F}$ *in the space spanned by* $\{\mathbf{u}_\ell\}_{\ell=1}^m$ *is given by*

$$\mathbf{F} := \mathbf{U}(\mathbf{P}^T\mathbf{U})^{-1}\mathbf{P}^T\mathbf{F}, \tag{3.3}$$

where $\mathbf{P} = [\mathbf{e}_{\varrho_1}, \ldots, \mathbf{e}_{\varrho_m}] \in \mathbb{R}^{n\times m}$*, and* $\mathbf{e}_{\varrho_i} = [0, ..0, \underbrace{1}_{\varrho_i}, 0, .., 0]^T \in \mathbb{R}^n$*,* $i = 1, .., m$*. The interpolation indices* $\{\varrho_1, \ldots, \varrho_m\}$ *are selected inductively from the basis* $\{\mathbf{u}_i\}_{i=1}^m$ *by the* DEIM *algorithm described below.*

Algorithm DEIM:
INPUT: $\{\mathbf{u}_\ell\}_{\ell=1}^m \subset \mathbb{R}^n$ linearly independent
OUTPUT: $\vec{\varrho} = [\varrho_1, \ldots, \varrho_m]^T \in \mathbb{R}^m$

1. $[|\rho|\ \varrho_1] = \max\{|\mathbf{u}_1|\}$
2. $\mathbf{U} = [\mathbf{u}_1],\ \mathbf{P} = [\mathbf{e}_{\varrho_1}],\ \vec{\varrho} = [\varrho_1]$
3. **for** $\ell = 2$ to m **do**
4. Solve $(\mathbf{P}^T\mathbf{U})\mathbf{c} = \mathbf{P}^T\mathbf{u}_\ell$ for $\mathbf{c}$
5. $\mathbf{r} = \mathbf{u}_\ell - \mathbf{U}\mathbf{c}$
6. $[|\rho|\ \varrho_\ell] = \max\{|\mathbf{r}|\}$
7. $\mathbf{U} \leftarrow [\mathbf{U}\ \mathbf{u}_\ell]$, $\mathbf{P} \leftarrow [\mathbf{P}\ \mathbf{e}_{\varrho_\ell}]$, $\vec{\varrho} \leftarrow \begin{bmatrix} \vec{\varrho} \\ \varrho_\ell \end{bmatrix}$
8. **end for**

Usually, the input basis $\mathbf{U}$ is obtained via POD method applied to nonlinear snapshots and inside the above algorithm we use $\mathbf{U}$ to denote some of its columns. This is motivated by the fact that the columns are added incrementally at each step, and once the algorithm reaches the finishing state, $\mathbf{U}$ is consistent with the initial notation proposed in Definition 1.

In the Algorithm DEIM we denoted by "max" the built-in Matlab function *max* with the same significance. Thus, this function applied at Step 6

by $[|\rho|\ \varrho_\ell] = \max\{|\mathbf{r}|\}$ leads to $|\rho| = |r_{\varrho_\ell}| = \max_{i=1,\ldots,n}\{|r_i|\}$, with the smallest index taken when the values along $|\mathbf{r}|$ contain more than one maximal element. Precisely, the index of the first one is returned. According to this algorithm, the DEIM procedure generates a set of indices inductively on the input basis in such a way that, at each iteration, the current selected index captures the maximum variation of the input basis vectors. The vector $\mathbf{r}$ can be viewed as the error between the input basis $\{\mathbf{u}_\ell\}_{\ell=1}^m$ and its approximation $\mathbf{Uc}$ from interpolating the basis $\{\mathbf{u}_\ell\}_{\ell=1}^{m-1}$ at the indices $\varrho_1, \ldots, \varrho_{m-1}$. The linear independence of the input basis $\{\mathbf{u}_\ell\}_{\ell=1}^m$ guarantees that, at each iteration, $\mathbf{r}$ is a nonzero vector and the output indices $\varrho_1, \ldots, \varrho_m$ are not repeating.

An error result for DEIM approximation of a nonlinear vector-valued function $\mathbf{F}$ is available in [7, Lemma 3.2], where the bound is obtained by limiting the local growth of a certain magnification factor. It was proved that $\mathbf{P}^T\mathbf{U}$ is always nonsingular and the greedy based DEIM selection process can be viewed in terms of minimizing the condition number of $\mathbf{P}^T\mathbf{U}$. Moreover, it was shown in [8, Theorem 3.1] that the error bounds in $2-$norm of the difference between the solutions of a full-order general nonlinear order differential equation and its corresponding POD-DEIM reduced order version can be approximated by the sums of the singular values corresponding to the neglected POD bases vectors of the state variables and nonlinear terms.

The POD and POD-DEIM reduced order models of the system (2.2) were developed by using a Galerkin projection and the techniques presented in this section.

4 Numerical results

The system (2.2) was solved numerically using a finite difference discretization. Let $0 = x_0 < x_1 < \cdots < x_n < x_{n+1} = 1$ be equally spaced points on the x-axis for generating the grid points on the dimensionless domain $\Omega = [0, 1]$, and take time domain $[0, T] = [0, 1]$. The corresponding spatial finite difference discretized system of (2.2) becomes a system of nonlinear ODEs. The semi-implicit Euler scheme was used to solve the discretized system of full dimension and POD and POD-DEIM reduced order systems.

The parameters were set to the following values ([20]): $D_N = D_M =$ 8.e-8, $D_P = D_Q =$ 7.5e-7, $D_K =$ 1.25e-6, $\chi_P =$ 1.5e-5, $\chi_Q =$ 1.5e-5, $\rho_1 = 2.5$, $\rho_2 = 0.25$, $\rho_3 = 2.5$, $\gamma_1 = 0.8$, $\gamma_2 = 0.01$, $\gamma_3 = 1$, $s_1 = 0.8$,

$s_2 = 0.2$, $s_3 = 0.8$, $c_1 = 0.3$, $c_2 = 0.004$, $c_3 = 0.2$, $\gamma_1 = 0.2$, $\gamma_2 = 0.1$ and $\gamma_3 = 0.01$. In our simulations we used the following initial conditions:

$$\begin{aligned} N_0(x) &= x(1-x)[0.75e^{-100(x-0.5)^2} + 0.25e^{-100(x-0.15)^2}], \\ M_0(x) &= x(1-x)[0.15e^{-100(x-0.35)^2} + 0.65e^{-100(x-0.5)^2}], \\ P_0(x) &= x(1-x)[0.075e^{-100(x-0.25)^2} + 0.075e^{-125(x-0.75)^2}], \\ P_0(x) &= x(1-x)[0.075e^{-125(x-0.15)^2} + 0.095e^{-175(x-0.65)^2}], \end{aligned}$$

and $K_0(x) = 0$. The number of spatial inner grid points on the x-axis was successively taken as 32, 64, 128, ..., 2048. The solution components of the problem (2.2) for a space configuration with 2048 internal nodes of each discretized variable are depicted in Figs. 1,2. Tables 1–4 and Figs. 3–5 show a significant improvement in computational time of the POD-DEIM reduced system compared to the POD reduced and the full-order system. Precisely, POD-DEIM reduces the computational time by a factor of $\mathcal{O}(10^2)$. The CPU time used in computing POD reduced system clearly reflects the dependency on the dimension of the original full-order system.

5 Conclusions

The model reduction technique combining POD with DEIM has been de-monstrated to be efficient for capturing the spatio-temporal dynamics of a multi-species host-parasitoid system with substantial reduction in both dimension and computational time by a factor of $\mathcal{O}(10^2)$. The failure to decrease complexity with the standard POD technique was clearly demonstrated by the comparative computational times shown in Tables 1–4 and Figs 3–5. DEIM was shown to be very effective in overcoming the deficiencies of POD with respect to the nonlinearities in the model under study. In order to increase the efficiency of the POD-DEIM approximation, a possible extension is to incorporate the POD-DEIM approach with higher-order FD schemes to improve the overall accuracy, especially due to the spatio-temporal heterogeneity and chemotaxis driven instability.

It is also interesting to compare the Discrete Empirical Interpolation Method with Gappy POD and Missing Point Estimation methods in a proper orthogonal decomposition framework applied to a higher order finite difference parasitoid model. The gappy POD procedure uses a POD basis to reconstruct missing, or "gappy" data and it was developed in [12]. The Missing Point Estimation method ([2]) relies on gappy POD technique and the reduced order model computes the Galerkin projections over a restricted subset of the spatial domain.

Table 1: CPU time of full-order system, POD and POD-DEIM reduced systems.

Internal Nodes n	CPU Time Full Dim	CPU Time POD	CPU Time POD–DEIM
32	5.407969e+00	5.317957e+00	1.715911e-01
64	5.254361e+00	5.347111e+00	1.680101e-01
128	5.607438e+00	5.710571e+00	1.696068e-01
256	6.847215e+00	6.614301e+00	1.809442e-01
512	8.610269e+00	7.600184e+00	2.016218e-01
1024	1.337721e+01	9.417793e+00	1.835292e-01
2048	2.653383e+01	1.312482e+01	1.812312e-01

Table 2: POD and POD-DEIM average relative errors for the components N and M – host species.

Internal Nodes n	$Error^{rel}$ POD – N	$Error^{rel}$ POD–DEIM – N	$Error^{rel}$ POD – M	$Error^{rel}$ POD–DEIM – M
32	3.482843e-14	3.516461e-14	1.645643e-13	1.657210e-13
64	1.388416e-14	1.414009e-14	9.331344e-14	9.348847e-14
128	1.653464e-14	1.661955e-14	7.420785e-14	7.175778e-14
256	4.718024e-15	4.669319e-15	1.590888e-14	1.634144e-14
512	2.736167e-14	2.732722e-14	1.716102e-14	2.124873e-14
1024	2.993938e-14	3.012212e-14	1.859783e-14	3.643836e-14
2048	9.590961e-15	1.042055e-14	4.956752e-14	1.216911e-13

Table 3: POD and POD-DEIM average relative errors for the components P and Q – parasitoid species.

Internal Nodes n	$Error^{rel}$ POD – P	$Error^{rel}$ POD–DEIM – P	$Error^{rel}$ POD – Q	$Error^{rel}$ POD–DEIM – Q
32	2.460205e-14	2.459944e-14	1.961488e-14	1.961738e-14
64	6.814060e-14	6.817415e-14	3.010484e-14	2.997920e-14
128	8.805397e-15	8.808601e-15	2.347853e-14	2.483260e-14
256	8.218387e-15	8.221235e-15	3.326519e-14	3.230054e-14
512	6.303037e-15	6.304210e-15	4.516320e-15	4.445458e-15
1024	1.758562e-14	1.720852e-14	3.067915e-15	3.980249e-15
2048	5.855724e-15	9.105957e-15	1.085525e-14	1.340351e-14

Table 4: POD and POD-DEIM average relative errors for the component K – chemoattractant.

Internal Nodes n	$Error^{rel}$ POD – K	$Error^{rel}$ POD–DEIM – K
32	5.987349e-14	6.004292e-14
64	3.937026e-14	3.981359e-14
128	3.118464e-14	3.054254e-14
256	1.440359e-14	1.604336e-14
512	3.286988e-14	3.330396e-14
1024	1.140597e-14	1.642880e-14
2048	2.154869e-14	4.431176e-14

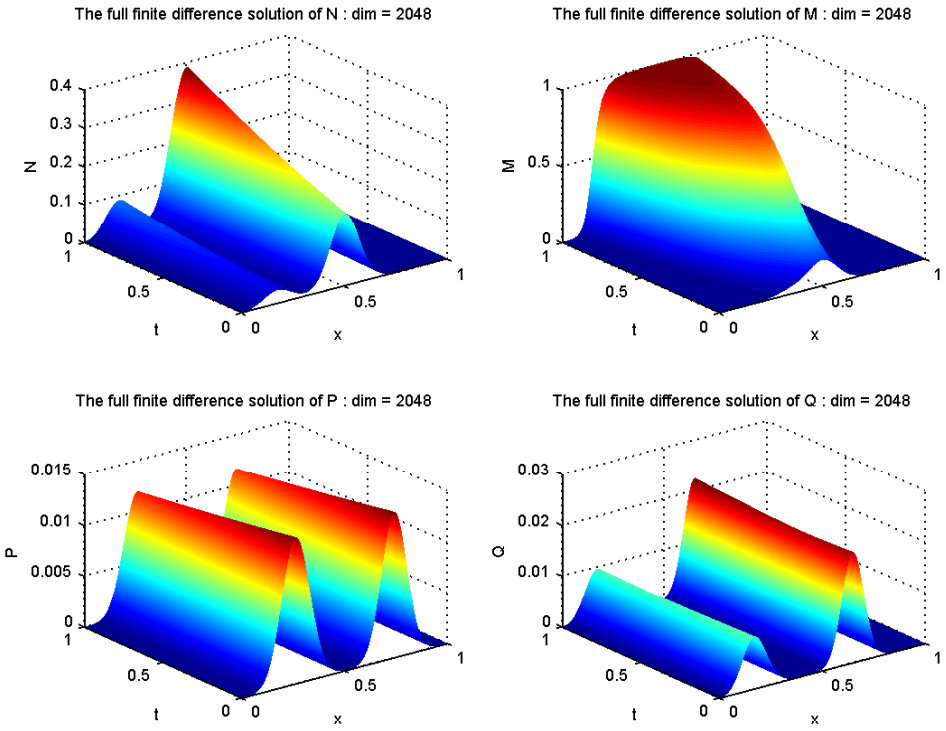

Figure 1: Solution plots (N, M, P, Q) of the model from the full-order system $(n = 2048)$.

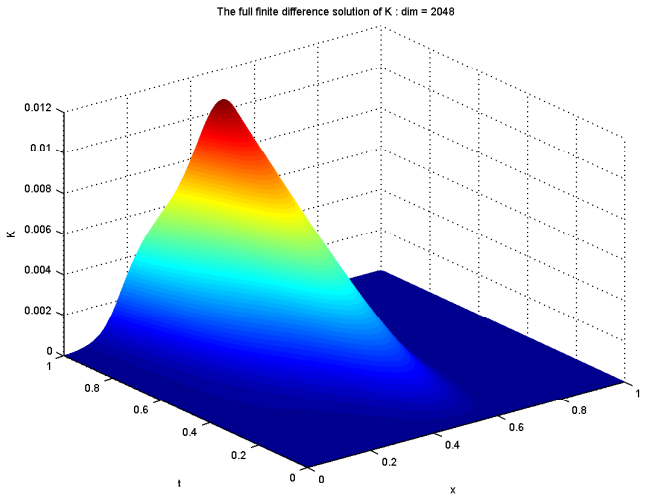

Figure 2: Solution plot K from the full-order system ($n = 2048$).

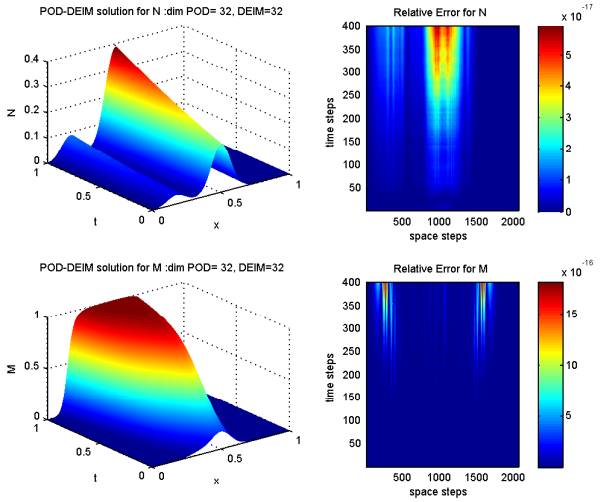

Figure 3: Solution plots (N, M) of the model from POD-DEIM reduced system (dimPOD=dimDEIM=32), with the corresponding average relative errors at the inner grid points ($n = 2048$).

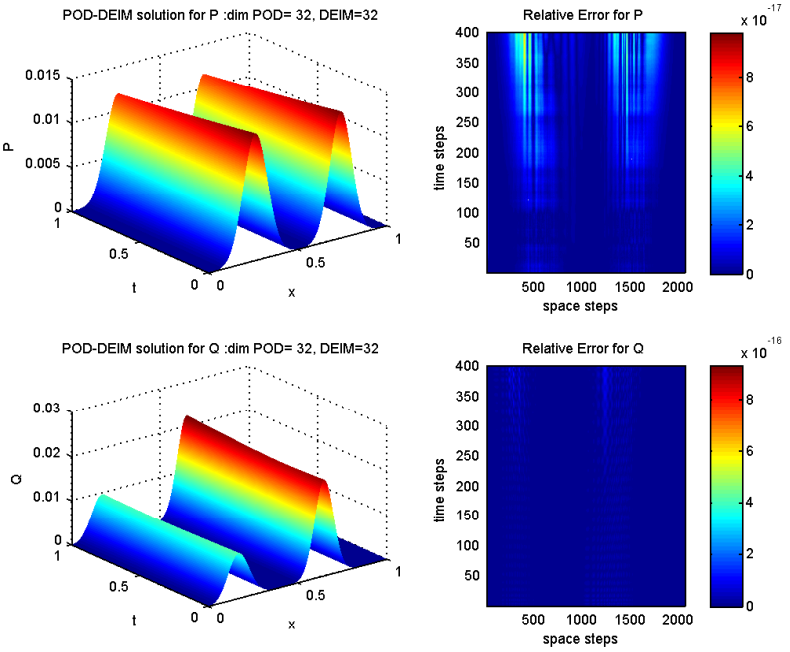

Figure 4: Solution plots (P, Q) of the model from POD-DEIM reduced system (dimPOD=dimDEIM=32), with the corresponding average relative errors at the inner grid points ($n = 2048$).

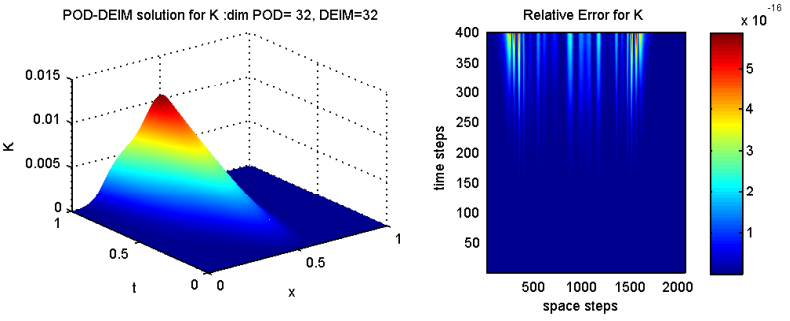

Figure 5: Solution plots K of the model from POD-DEIM reduced system (dimPOD=dimDEIM=32), with the corresponding average relative errors at the inner grid points ($n = 2048$).

Acknowledgement. The first author acknowledges the support of the grant of the Romanian National Authority for Scientific Research, CNCS - UEFISCDI, project number PN-II-ID-PCE-2011-3-0563, contract no. 343/ 5.10.2011 "Models from medicine and biology: mathematical and numerical insights". Prof. I.M. Navon acknowledges the support of NSF grant ATM-0931198.

References

[1] A.C. Antoulas. *Approximation of Large-Scale Dynamical Systems.* Advances in Design and Control. SIAM, Philadelphia, 2005.

[2] P. Astrid, S. Weiland, K. Willcox, T. Backx. Missing point estimation in models described by proper orthogonal decomposition. *IEEE Trans. Automat. Control.* 53(10):2237–2251, 2008.

[3] J.A. Atwell, B.B. King. Proper orthogonal decomposition for reduced basis feedback controllers for parabolic equations. *Math. Comput. Modelling.* 33(1-3):1–19, 2001.

[4] M. Barrault, Y. Maday, N.C. Nguyen, A.T. Patera. An "empirical interpolation" method: application to efficient reduced-basis discretization of partial differential equations. *C.R. Math. Acad. Sci. Paris.* 339(9):667–672, 2004.

[5] G. Berkooz, P. Holmes, J. Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Ann. Rev. Fluid Mech.* 25:777–786, 1993.

[6] S. Chaturantabut. Dimension Reduction for Unsteady Nonlinear Partial Differential Equations via Empirical Interpolation Methods. Technical Report TR09-38, CAAM, Rice University, 2008.

[7] S. Chaturantabut, D.C. Sorensen. Nonlinear model reduction via discrete empirical interpolation. *SIAM J. Sci. Comput.* 32(5):2737–2764, 2010.

[8] S. Chaturantabut, D.C. Sorensen. A state space error estimate for POD-DEIM nonlinear model reduction. *SIAM J. Numer. Anal.* 50(1):46–63, 2012.

[9] G. Dimitriu, N. Apreutesei. Comparative study with data assimilation experiments using proper orthogonal decomposition method. *Lecture Notes in Comput. Sci.* 4818:393–400, 2008.

[10] G. Dimitriu, N. Apreutesei, R. Ştefănescu. Numerical simulations with data assimilation using an adaptive POD procedure. *Lecture Notes in Comput. Sci.* 5910:165–172, 2010.

[11] G. Dimitriu, I.M. Navon, R. Ştefănescu. Application of POD-DEIM Approach for Dimension Reduction of a Diffusive Predator-Prey System with Allee Effect. *Lecture Notes in Comput. Sci.* 8353:373–381, 2014.

[12] R. Everson, L. Sirovich. Karhunen-Loève procedure for gappy data. *J. Opt. Soc. Am.* A 12:1657–1664, 1995.

[13] M.A. Grepl, Y. Maday, N.C. Nguyen, A.T. Patera. Efficient reduced-basis treatment nonaffine and nonlinear partial differential equations. *ESAIM Modélisation Mathématique et Analyse Numérique.* 41(3):575–605, 2007.

[14] M. Gubisch and S. Volkwein. Proper Orthogonal Decomposition for Linear-Quadratic Optimal Control. *University of Konstanz, Technical report*, 2013.

[15] K. Kunisch, S. Volkwein. Control of the Burgers' equation by a reduced order approach using proper orthogonal decomposition. *J. Optim. Theory Appl.* 102(2):345–371, 1999.

[16] K. Kunisch, S. Volkwein, and L. Xie. HJB-POD-Based Feedback Design for the Optimal Control of Evolution Problems. *SIAM J. Appl. Dyn. Syst*, 3(4):701–722, 2004.

[17] O. Lass, S. Volkwein. POD Galerkin schemes for nonlinear elliptic-parabolic systems. *Konstanzer Schriften in Mathematik.* 301:1430–3558, 2012.

[18] Y. Maday, N.C. Nguyen, A.T. Patera, G.S.H. Pau. A General Multipurpose Interpolation Procedure: the Magic Points. *Commun. Pure Appl. Anal.* 8(1):383–404, 2009.

[19] I.G. Pearce, M.A.J. Chaplain, P.G. Schofield, A.R.A. Anderson, S.F. Hubbard. Modelling the spatio-temporal dynamics of multi-species host-parasitoid interactions: heterogeneous patterns and ecological implications. *J. Theor. Biol.* 241:876–886, 2006.

[20] I.G. Pearce, M.A.J. Chaplain, P.G. Schofield, A.R.A. Anderson, S.F. Hubbard. Chemotaxis-induced spation-temporal heterogeneity in multi-species host-parasitoid systems. *J. Math. Biol.* 55:365–388.

[21] L. Sirovich. Turbulence and the dynamics of coherent structures. I. Coherent structures. *Quart. Appl. Math.* 45(3):561–571, 1987.

[22] L. Sirovich. Turbulence and the dynamics of coherent structures. II. Symmetries and transformations. *Quart. Appl. Math.* 45(3):573–582, 1987.

[23] L. Sirovich. Turbulence and the dynamics of coherent structures. III. Dynamics and scaling. *Quart. Appl. Math.* 45(3):583–590, 1987.

[24] R. Ştefănescu, I.M. Navon. POD/DEIM nonlinear model order reduction of an ADI implicit shallow water equations model. *J. Comput. Phys.* 237:95–114, 2013.

[25] R. Ştefănescu, A. Sandu, I.M. Navon. Comparison of POD reduced order strategies for the nonlinear 2D shallow water equations. *Internat. J. Numer. Methods Fluids* 76(8):497–521, 2014.

[26] R. Ştefănescu, A. Sandu. Efficient Approximation of Sparse Jacobians for Time-Implicit Reduced Order Models. *Virginia Polytechnic Institute and State University, Technical Report*, TR 15, 2014.

[27] T. Tonn. Reduced-Basis Method (RBM) for Non-Affine Elliptic Parametrized PDEs. *Dissertation, Ulm University*, 2012.

[28] D. Wirtz, D. C. Sorensen, B. Haasdonk. A Posteriori Error Estimation for DEIM Reduced Nonlinear Dynamical Systems. *SIAM J. Sci. Comput.*, 36(2):A311–A338, 2014.

Ann. Acad. Rom. Sci.
Ser. Math. Appl.
ISSN 2066-6594 Vol. 7, No. 1/2015

ON THE NUMERICAL APPROXIMATION OF THE NONLINEAR PHASE-FIELD EQUATION SUPPLIED WITH NON-HOMOGENEOUS DYNAMIC BOUNDARY CONDITIONS. CASE 1D*

Costică Moroşanu†

Dedicated to the memory of Prof. Dr. Viorel Arnăutu

Abstract

The paper is concerned with the numerical analysis of a scheme of fractional steps type, associated to the nonlinear phase-field (Allen-Cahn) equation, endowed with non-homogeneous dynamic boundary conditions (depending both on the time and space variables). To approximate the solution of the linear parabolic equation, introduced by such approximating schemes, a first-order **IM**plicit **B**ackward **D**ifferentiation **F**ormula (**1-IMBDF**) is considered. A conceptual numerical algorithm and numerical experiments in one dimension are performed too.

MSC: 35K55, 65N06, 65N12, 65YXX, 80AXX

keywords: Boundary value problems for nonlinear parabolic PDE, fractional steps method, convergence of numerical method, computer aspects of numerical algorithm, phase-changes, dynamic boundary conditions.

*Accepted for publication in revised form on February 4-th, 2015

†costica.morosanu@uaic.ro "Al. I. Cuza" University, Iaşi, 700506, Romania

1 Introduction

Consider the following nonlinear parabolic boundary value problem with respect to the unknown function φ:

$$\begin{cases} \alpha\xi\dfrac{\partial}{\partial t}\varphi - \xi\Delta\varphi = \dfrac{1}{2\xi}(\varphi-\varphi^3) & \text{in } \ Q=[0,T]\times\Omega \\ \xi\dfrac{\partial}{\partial \nu}\varphi + \alpha\xi\dfrac{\partial}{\partial t}\varphi - \Delta_\Gamma\varphi + c_0\varphi = w(t,x) & \text{on } \ \Sigma=[0,T]\times\partial\Omega \\ \varphi(0,x)=\varphi_0(x) & \text{on } \ \Omega, \end{cases} \tag{1.1}$$

where:

- Ω is a bounded domain in $\mathbb{R}$ with boundary $\partial\Omega=\Gamma$ and $T>0$ stands for some final time;
- $\varphi(t,x)$ is the *phase function* (used to distinguish between the states (phases) of a material which occupies the region Ω at every time $t\in[0,T]$);
- α (the *relaxation time*), ξ (the *measure of the interface thickness*) and c_0 are positive constants;
- Δ_Γ is the Laplace-Beltrami operator;
- $w(t,x)\in W_p^{1-\frac{1}{2p},2-\frac{1}{p}}(\Sigma)$ is a given function and p satisfies
$$p\geq\frac{3}{2}; \tag{1.2}$$
- $\varphi_0\in W_\infty^{2-\frac{2}{p}}(\Omega)$ verifying $\xi\frac{\partial}{\partial\nu}\varphi_0-\Delta_\Gamma\varphi_0+c_0\varphi_0=w(0,x)$ on Γ.

Equation $(1.1)_1$ was introduced initially by Allen and Cahn (see [1]) to describe the motion of anti-phase boundaries in crystalline solids. Actually, the Allen-Cahn model is widely applied to moving interface problems, such as the mixture of two incompressible fluids, the nucleation of solids, vesicle membranes, etc. Also, the nonlinear parabolic equation $(1.1)_1$ appears in the Caginalp's phase-field transition system (see [4]) describing the transition between the solid and liquid phases in the solidification process of a material occupying a region Ω (see [6]).

Following the strategy used in [5] and [9], the nonlinear parabolic boundary value problem (1.1) can be rewritten suitably in the following form:

$$\begin{cases} \alpha\xi\frac{\partial}{\partial t}\varphi - \xi\Delta\varphi = \frac{1}{2\xi}(\varphi-\varphi^3) & \text{in } Q \\ \varphi = \psi & \text{on } \Sigma \\ \xi\frac{\partial}{\partial \nu}\varphi + \alpha\xi\frac{\partial}{\partial t}\psi - \Delta_\Gamma\psi + c_0\psi = w(t,x) & \text{on } \Sigma \\ \varphi(0,x) = \varphi_0(x) & x\in\Omega \\ \psi(0,x) = \psi_0(x) & x\in\Gamma, \end{cases} \tag{1.3}$$

where the new variable $\psi = \varphi$, $\psi(0,x) = \varphi_0$ on Γ, is introduced in order to treat the dynamic boundary conditions $(1.1)_2$ as a parabolic equation for ψ on the boundary Γ, with $\psi_0 \in W_\infty^{2-\frac{2}{p}}(\Gamma)$, $\varphi_0 = \psi_0$ on Γ and, for the remaining data in (1.1), we keep the meanings already formulated.

As regards the existence in (1.3), it is known that under appropriate conditions on φ_0 and w, there exists a unique solution $(\varphi,\psi) \in W_p^{1,2}(Q) \times W_p^{1,2}(\Sigma)$, $p \geq \frac{3}{2}$ (see [5, Theorem 2.1]). Here we have used the standard notation for Sobolev spaces, namely, given a positive integer k and $1 \leq p \leq \infty$, we denote by $W_p^{k,2k}(Q)$ the usual Sobolev space on Q:

$$W_p^{k,2k}(Q) = \left\{ y \in L^p(Q) : \ \frac{\partial^r}{\partial t^r}\frac{\partial^s}{\partial x^s}\, y \in L^p(Q), \text{ for } 2r+s \leq k \right\},$$

i.e., the space of functions whose t-derivatives and x-derivatives up to the order k and $2k$, respectively, belong to $L^p(Q)$. Also, we have used the Sobolev spaces $W_p^l(\Omega)$, $W_p^{\frac{l}{2},l}(\Sigma)$ with nonintegral l for the initial and boundary conditions, respectively (see [7, Chapter 1] and references therein).

Numerical investigation of the nonlinear parabolic problem (1.1), subject to various other types of boundary conditions, have been made in [2], [3], [7] and [8]. The main novelty of this work is the presence of the nonhomogeneous dynamic boundary conditions $(1.1)_2$, untreated numerically until now (to our knowledge) in the mathematical literature and which makes the present nonlinear parabolic problem (1.1) to be more accurate in describing many important phenomena of two-phase systems: *superheating*, *supercooling*, *the effects of surface tension*, *separating zones*, etc; in particular, the interactions with the walls in confined systems. Consequently, a wide variety of industrial applications are covered.

In order to approximate the solution of the nonlinear boundary value problem (1.3) (in fact, the solution of problem (1.1)), a *scheme of fractional steps type* was introduced and analyzed in [9], namely, for every $\varepsilon > 0$, it was associated to problem (1.3) the following approximating scheme (see also [2-3], [6-8]):

$$\begin{cases} \alpha\xi \frac{\partial}{\partial t}\varphi^\varepsilon - \xi\Delta\varphi^\varepsilon = \frac{1}{2\xi}\varphi^\varepsilon & \text{in } Q_i^\varepsilon \\ \xi \frac{\partial}{\partial \nu}\varphi^\varepsilon + \alpha\xi \frac{\partial}{\partial t}\psi^\varepsilon - \Delta_\Gamma \psi^\varepsilon + c_0\psi^\varepsilon = w(t,x) & \text{on } \Sigma_i^\varepsilon \\ \varphi^\varepsilon(i\varepsilon, x) = z(\varepsilon, \varphi_-^\varepsilon(i\varepsilon, x)) & \text{on } \Omega \\ \psi^\varepsilon(i\varepsilon, x) = \varphi^\varepsilon(i\varepsilon, x) & \text{on } \Gamma, \end{cases} \tag{1.4}$$

where $Q_i^\varepsilon = [i\varepsilon, (i+1)\varepsilon] \times \Omega$, $\Sigma_i^\varepsilon = [i\varepsilon, (i+1)\varepsilon] \times \Gamma$ and $z(\varepsilon, \varphi_-^\varepsilon(i\varepsilon, x))$ is the solution of the Cauchy problem:

$$\begin{cases} z'(s) + \frac{1}{2\xi} z^3(s) = 0 & s \in [0, \varepsilon] \\ z(0) = \varphi_-^\varepsilon(i\varepsilon, x) & \text{on} \quad \Omega \\ \varphi_-^\varepsilon(0, x) = \varphi_0(x) & \text{on} \quad \Omega \\ \varphi_-^\varepsilon(0, x) = \psi_0(x) & \text{on} \quad \Gamma, \end{cases} \tag{1.5}$$

for $i = 0, 1, \cdots, M_\varepsilon - 1$, with $M_\varepsilon = \left[\frac{T}{\varepsilon}\right]$, $Q_{M_\varepsilon - 1}^\varepsilon = [(M_\varepsilon - 1)\varepsilon, T] \times \Omega$, $\Sigma_{M_\varepsilon - 1}^\varepsilon = [(M_\varepsilon - 1)\varepsilon, T] \times \Gamma$ and φ_-^ε stands for the left-hand limit of φ^ε.

In other words, the fractional steps method consists in decoupling the nonlinear problem (1.3) in a linear parabolic boundary value problem, expressed on a partition of the time interval $[0, T]$ (composed from M_ε subintervals, the first $M_\varepsilon - 1$ having the same length ε) and a nonlinear ordinary differential equation containing the nonlinearity φ^3. Accordingly, the advantage of this approach consists in simplifying the numerical computation of the process of approximation for the solution of nonlinear problem (1.1).

Invoking again the Theorem 2.1 in [5], we have that there is a unique solution to (1.4)-(1.5), namely: $(\varphi^\varepsilon, \psi^\varepsilon) \in W_p^{1,2}(Q_i^\varepsilon) \times W_p^{1,2}(\Sigma_i^\varepsilon)$, with $p \geq \frac{3}{2}$ and $i = 0, 1, \cdots, M_\varepsilon - 1$.

Owing to the Lions and Peetre embedding theorem, we know that $W_p^{1,2}(Q) \subset L^\infty(Q)$ if $p \geq \frac{3}{2}$ (see [7, Chapter 1] and references therein) and thus, for later use, we will introduce the sets:
$W_Q = L^2([0,T]; H^1(\Omega)) \cap L^\infty(Q)$ and $W_\Sigma = L^2([0,T]; H^1(\Gamma)) \cap L^\infty(\Sigma)$.

Definition 1 *By a weak solution of the nonlinear problem* (1.3) *we mean a pair of functions* $(\varphi, \psi) \in W_Q \times W_\Sigma$, $\varphi = \psi$ *on* Σ, *which satisfies* (1.3) *in the following sense:*

$$\alpha\xi \int_Q \left(\frac{\partial}{\partial t}\varphi, \phi_1\right) dt\, dx + \xi \int_Q \nabla\varphi \nabla\phi_1\, dt\, dx \tag{1.6}$$

$$+\alpha\xi \int_\Sigma \left(\frac{\partial}{\partial t}\psi, \phi_2\right) dt\, d\gamma + \int_\Sigma \nabla\psi \nabla\phi_2\, dt\, d\gamma + c_0 \int_\Sigma \psi\phi_2\, dt d\gamma$$

$$= \frac{1}{2\xi} \int_Q (\varphi - \varphi^3)\phi_1\, dt\, dx + \int_\Sigma w\phi_2\, dt d\gamma$$

$\forall (\phi_1, \phi_2) \in L^2([0,T]; H^1(\Omega)) \times L^2([0,T]; H^1(\Gamma))$, *and* $\varphi(0,x) = \varphi_0(x)$ in Ω.

The symbols $\int_Q$ and $\int_\Sigma$ above denote the duality between $L^2([0,T]; H^1(\Omega))$ and $L^2([0,T]; H^1(\Omega)')$, as well as $L^2([0,T]; H^1(\Gamma))$ and $L^2([0,T]; H^1(\Gamma)')$, respectively.

The following result (see [3], [7]) establishes the relationship between the solution (φ, ψ) in (1.3) and the solution $(\varphi^\epsilon, \psi^\epsilon)$ in (1.4)-(1.5).

Theorem 1 *Assume that* $\varphi_0(x) \in W_\infty^{2-\frac{2}{q}}(\Omega)$, *satisfying* $\xi\frac{\partial}{\partial\nu}\varphi_0 - \Delta_\Gamma\varphi_0 + c_0\varphi_0 = w(0,x)$ *on* Γ, *and* $w(t,x) \in W_p^{1-\frac{1}{2p}, 2-\frac{1}{p}}(\Sigma)$. *Let* $(\varphi^\varepsilon, \psi^\varepsilon)$ *be the solution of the approximating scheme* (1.4)-(1.5). *Then for* $\varepsilon \to 0$, *one has*

$$(\varphi^\varepsilon, \psi^\varepsilon) \to (\varphi^\star, \psi^\star) \quad \text{strongly in} \quad L^2(\Omega) \times L^2(\Gamma) \quad \text{for any} \quad t \in (0,T], \tag{1.8}$$

where $(\varphi^\star, \psi^\star) \in W_Q \times W_\Sigma$ *is the weak solution of the nonlinear equation* (1.3).

The outline of the paper is as follows: in Section 2 we have introduced the discrete equations corresponding to (1.4)-(1.5); consequently, a conceptual numerical algorithm has been formulated: **Alg_1-IMBDF_dbc**. A stability result for this approach is stated and proved in the next Section. Some numerical experiments are reported in the last Section.

2 Numerical method

In this section we are concerned with the numerical approximation of the solution $(\varphi^\varepsilon, \psi^\varepsilon)$ to (1.4)-(1.5). As already stated, we will work in one dimension and then $\Delta\varphi^\varepsilon = \varphi^\varepsilon_{xx}$, $\Delta_\Gamma\psi^\varepsilon = \psi^\varepsilon_{xx}$ and $\frac{\partial}{\partial\nu}\varphi^\varepsilon = \frac{\partial}{\partial x}\varphi^\varepsilon \cdot \nu = \mp\varphi^\varepsilon_x$ (i.e., (see [7, Chapter 1, p. 27]), the directional derivative of φ^ε in the direction of the outward pointing unit normal vector ν).

Let $\Omega = (0, b) \subset I\!R_+$ and we introduce over it the grid with N equidistant nodes

$$x_j = (j-1)dx \qquad j = 1, 2, \ldots, N, \quad dx = \frac{b}{N-1}.$$

Accordingly, the boundary Γ is given by the set of points $\{x_1 = 0, x_N = b\}$.

Considering $M \equiv M_\varepsilon$ as the number of equidistant nodes in which is divided the time interval $[0, T]$, we set

$$t_i = (i-1)\varepsilon \qquad i = 1, 2, \ldots, M, \quad \varepsilon = \frac{T}{M-1}.$$

We denote by φ^i_j the approximate values in the point (t_i, x_j) of the unknown function φ^ε. More precisely

$$\varphi^i_j = \varphi^\varepsilon(t_i, x_j) \qquad i = 1, 2, \ldots, M, \quad j = 1, 2, \ldots, N,$$

i.e., for the later use

$$\varphi^i \overset{\text{not}}{=} \left(\varphi^i_1, \varphi^i_2, \ldots, \varphi^i_N\right)^T \qquad i = 1, 2, \ldots, M. \tag{2.1}$$

We continue by explaining how we will treat each term from (1.4)-(1.5). Owing to the relation $(1.4)_4$ and knowing that $\Gamma = \{x_1, x_N\}$, we can put

$$\begin{cases} \psi^i_1 = \psi^\varepsilon(t_i, x_1) = \varphi^\varepsilon(t_i, x_1) = \varphi^i_1 \\ \\ \psi^i_N = \psi^\varepsilon(t_i, x_N) = \varphi^\varepsilon(t_i, x_N) = \varphi^i_N \end{cases} \qquad i = 1, 2, \ldots, M. \tag{2.2}$$

The Laplace operator in $(1.4)_1$ will be approximated by a *second order centred finite differences*, that is, for $i = 1, 2, ..., M$:

$$\varphi^\varepsilon_{xx}(t_i, x_j) = \Delta_{dx}\varphi^i_j \approx \frac{\varphi^i_{j-1} - 2\varphi^i_j + \varphi^i_{j+1}}{dx^2} \qquad j = 1, 2, ..., N, \tag{2.3}$$

where Δ_{dx} is the discrete Laplace operator, depending on the step-size dx. Corresponding to the Laplace-Beltrami operator in $(1.4)_2$, we will use the

same approximating scheme as above, which, correlated with $(1.4)_4$ and (2.2), gives us

$$\begin{cases} \psi^{\varepsilon}_{xx}(t_i, x_1) = \Delta_{dx}\psi^i_1 \approx \frac{\varphi^i_0 - 2\varphi^i_1 + \varphi^i_2}{dx^2} \\ \\ \psi^{\varepsilon}_{xx}(t_i, x_N) = \Delta_{dx}\psi^i_N \approx \frac{\varphi^i_{N-1} - 2\varphi^i_N + \varphi^i_{N+1}}{dx^2} \end{cases} \quad i = 1, 2, ..., M, \tag{2.4}$$

where φ^i_0 and φ^i_{N+1} are dummy variables.

Involving the separation of variables method to solve the Cauchy problem (1.5) (see [2], [6], [7], [8]), we obtain

$$\begin{cases} z(\varepsilon, \varphi^{\varepsilon}_-(t_1, x)) = z(\varepsilon, \varphi_0(x)) = \varphi_0(x)\sqrt{\frac{\xi}{\xi + \varepsilon\varphi_0(x)}}, \\ \\ z(\varepsilon, \varphi^{\varepsilon}_-(t_i, x)) = \varphi^{\varepsilon}_-(t_i, x)\sqrt{\frac{\xi}{\xi + \varepsilon\varphi^{\varepsilon}_-(t_i, x)}} \end{cases} \quad i = 2, ..., M-1. \tag{2.5}$$

Remembering that $\partial\Omega = \Gamma = \{x_1, x_N\}$, the boundary conditions $(1.4)_2$ can be rewritten as follows

$$\begin{cases} -\xi\varphi^{\varepsilon}_x(x_1) + \alpha\xi\frac{\partial}{\partial t}\psi^{\varepsilon}(t_i, x_1) - \psi^{\varepsilon}_{xx}(t_i, x_1) + c_0\psi^{\varepsilon}(t_i, x_1) = w(t_i, x_1) \\ \\ \xi\varphi^{\varepsilon}_x(x_N) + \alpha\xi\frac{\partial}{\partial t}\psi^{\varepsilon}(t_i, x_N) - \psi^{\varepsilon}_{xx}(t_i, x_N) + c_0\psi^{\varepsilon}(t_i, x_N) = w(t_i, x_N), \end{cases} \tag{2.6}$$

for $i = 1, 2, \ldots, M$, where the sign in the front of $\frac{\partial}{\partial\nu}\varphi^{\varepsilon} = \varphi^{\varepsilon}_x \cdot \nu$ is $-$ $(+)$ because the normal to $[0 = x_1, b = x_N]$ at x_1 (x_N) point is in the negative (positive) direction (i.e. the unit normal vector $\nu = \mp 1$ at 0 and b, respectively).

Now, using in (2.6) a forward (backward) finite differences to approximate $\varphi^{\varepsilon}_x(x_1)$ $(\varphi^{\varepsilon}_x(x_N))$ and, taking into account the relations (2.2) and (2.4), we get

$$\begin{cases} -\xi\frac{\varphi^i_2 - \varphi^i_1}{dx} + \alpha\xi\frac{\partial}{\partial t}\psi^{\varepsilon}(t_i, x_1) - \Delta_{dx}\psi^i_1 + c_0\psi^i_1 = w^i_1 \\ \\ \xi\frac{\varphi^i_N - \varphi^i_{N-1}}{dx} + \alpha\xi\frac{\partial}{\partial t}\psi^{\varepsilon}(t_i, x_N) - \Delta_{dx}\psi^i_N + c_0\psi^i_N = w^i_N, \end{cases} \tag{2.7}$$

where $w^i_1 = w(t_i, x_1)$, $w^i_N = w(t_i, x_N)$, $i = 1, 2, ..., M$.

For approximating the partial derivative with respect to time, we employed a *first-order scheme*, namely:

$$\begin{cases} \frac{\partial}{\partial t}\varphi^{\varepsilon}(t_i,x_j) \approx \frac{\varphi_j^i-\varphi_j^{i-1}}{\varepsilon} & i=2,3,\ldots,M, \quad j=1,2,\ldots,N \\ \\ \frac{\partial}{\partial t}\psi^{\varepsilon}(t_i,x_j) \approx \frac{\psi_j^i-\psi_j^{i-1}}{\varepsilon} & i=2,3,\ldots,M, \quad j\in\{1,N\}. \end{cases} \tag{2.8}$$

Finally we refer to the right hand in $(1.4)_1$ that is $\frac{1}{2\xi}\varphi^{\varepsilon}(t_i,x_j)$. To approximate this quantity (the *reaction term*), we will involve an implicit formula (see [8]), i.e.:

$$\frac{1}{2\xi}\varphi^{\varepsilon}(t_i,x_j) \approx \frac{1}{2\xi}\varphi_j^i \quad i=1,2,\ldots,M, \quad j=1,2,\ldots,N. \tag{2.9}$$

We are now ready to build the **1-IMBDF** approximating scheme. To do this, we begin by replacing in $(1.4)_1$ the approximations stated in (2.3), $(2.8)_1$ and (2.9). We deduce

$$\alpha\xi\ \frac{\varphi_j^i-\varphi_j^{i-1}}{\varepsilon} - \xi\Delta_{dx}\varphi_j^i = \frac{1}{2\xi}\varphi_j^i, \quad i=\overline{2,M}, \quad j=\overline{1,N}. \tag{2.10}$$

We continue by replacing in $(1.4)_2$ the approximations stated in (2.4), (2.7) and $(2.8)_2$ which leads to

$$\begin{cases} \alpha\xi\frac{\psi_1^i-\psi_1^{i-1}}{\varepsilon} - \xi\frac{\varphi_2^i-\varphi_1^i}{dx} - \Delta_{dx}\psi_1^i + c_0\psi_1^i = w_1^i, \\ \\ \alpha\xi\frac{\psi_N^i-\psi_N^{i-1}}{\varepsilon} + \xi\frac{\varphi_N^i-\varphi_{N-1}^i}{dx} - \Delta_{dx}\psi_N^i + c_0\psi_N^i = w_N^i, \quad i=\overline{2,M}. \end{cases} \tag{2.11}$$

Substituting in (2.10) and (2.11) the approximations of $\Delta_{dx}\varphi_j^i$, $\Delta_{dx}\psi_1^i$ and $\Delta_{dx}\psi_N^i$, expressed by (2.3) and (2.4), respectively, using (2.2) and arranging convenient, we obtain that (1.4) is discretized as follows

$$\begin{cases} -c_2\varphi_{j-1}^i + \left[c_1+2c_2-\frac{1}{2\xi}\right]\varphi_j^i - c_2\varphi_{j+1}^i = c_1\varphi_j^{i-1} \quad j=\overline{1,N}, \\ \\ \left[c_1+c_3+2+c_0\right]\varphi_1^i - (1+c_3)\varphi_2^i = w_1^i + c_1\varphi_1^{i-1} + \varphi_0^i, \\ \\ -(1+c_3)\varphi_{N-1}^i + \left[c_1+c_3+2+c_0\right]\varphi_N^i = w_N^i + c_1\varphi_N^{i-1} + \varphi_{N+1}^i, \end{cases} \tag{2.12}$$

for $i=2,3,\ldots,M$, where

$$c_1=\frac{\alpha\xi}{\varepsilon}, \quad c_2=\frac{\xi}{dx^2} \quad \text{and} \quad c_3=\frac{\xi}{dx}.$$

In order to compute the matrix $\left(\varphi_j^i\right)_{i=\overline{2,M},\ j=\overline{1,N}}$, the linear system (2.12) will be solved ascending with respect to the time levels. For the first time level ($i = 1$), the values of φ_j^1 are computed using $(1.4)_3$ and (2.5). For more details on implementing this computation process which involves the variable z, see the cycle "*For $i = 2$ to M do*" in the algorithm "**Alg_1-IMBDF_dbc**" listed below.

Moreover, let us point out from (2.12) that we have N unknowns for each time-level i, $i = 2, 3, ..., M$ (see and (2.1)).

If, corresponding to $j = 1$ and $j = N$ we take $\varphi_0^i = \varphi_1^i$ and $\varphi_{N+1}^i = \varphi_N^i$, than the system (2.12) can be rewritten in matrix form as

$$A\varphi^i = B\varphi^{i-1} + d^i \quad i = 2, 3, ..., M, \tag{2.13}$$

where

$$A = \begin{pmatrix} a_1 & -(1+c_2+c_3) & 0 & \cdots & 0 & 0 & 0 \\ -c_2 & c_1+2c_2-\frac{1}{2\xi} & -c_2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -c_2 & c_1+2c_2-\frac{1}{2\xi} & -c_2 \\ 0 & 0 & 0 & \cdots & 0 & -(1+c_2+c_3) & a_1 \end{pmatrix}$$

$a_1 = c_1 + 2c_2 + c_3 + c_0 + 1 - \frac{1}{2\xi}$,

$$\mathrm{B} = \begin{pmatrix} 2c_1 & 0 & \cdots & 0 & 0 \\ 0 & c_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & c_1 & 0 \\ 0 & 0 & \cdots & 0 & 2c_1 \end{pmatrix} \qquad \mathrm{d}^i = \begin{pmatrix} w_1^i \\ 0 \\ \vdots \\ 0 \\ w_N^i \end{pmatrix}.$$

Therefore, the general design of the algorithm to calculate the approximate solution to the nonlinear system (1.4)-(1.5), via *fractional steps method* and *1-IMBDF*, is the following one

`Begin` **Alg_1-IMBDF_dbc**
  `Choose` $T > 0$, $b > 0$;
  `Choose` $M > 0$, $N > 0$; `compute` ε `and` dx;
  `Choose` φ_0 `and` w;
  `Set` $\psi_0(x_1) = \varphi_0(x_1)$ `and` $\psi_0(x_N) = \varphi_0(x_N)$;
  `Compute` $\varphi_1^1 = \varphi_-^\epsilon(0, x_1) = \psi_0(x_1)$ `from` $(1.5)_4$;
  `For` $j = 2$ `to` $N - 1$ `do`
    `Compute` $\varphi_j^1 = \varphi_-^\epsilon(0, x_j) = \varphi_0(x_j)$ `from` $(1.5)_3$;

```
        End-for;
        Compute φ_N^1 = φ_-^ε(0, x_N) = ψ_0(x_N) from (1.5)_4;
        For i = 2 to M do
            Compute w_1^{i-1} and w_N^{i-1};
            Compute φ^{i-1} = z(ε, φ_-^ε(t_{i-1}, ·)) using (2.5);
            Compute φ^i solving the linear system (2.13);
        End-for;
    End.
```

As it is well known, most initial value problems reduce to solving large sparse linear systems of the form (2.13). For later use regarding the numerical implementation of the conceptual algorithms **Alg_1-IMBDF_dbc**, we proof the following

Lemma 1. *If*

$$c_1 + 2c_2 + c_3 + c_0 + 1 - \frac{1}{2\xi} \neq 0 \quad \text{and} \quad c_1 + 2c_2 - \frac{1}{2\xi} \neq 0, \tag{2.14}$$

then the matrix coefficients in linear system (2.13) *can be factored into the product of a lower-triangular matrix and an upper-triangular matrix* (*LU - factorization*).

Proof. Let denote by a_{mn}, $m, n = 1, 2, \cdots, N$, the elements of matrix coefficients in linear system (2.13). Analyzing the main diagonal elements of block matrices A, we first find that, owing to the hypothesis expressed by (2.14), second part, the coefficients a_{nn}, $n = 2, 3, \cdots, N-1 \neq 0$. Observing now that $a_1 \neq 0$ reflect the assumptions expressed in (2.14), first part, we find easily that $a_{nn} \neq 0 \ \forall n = 1, 2, \cdots, N$. So Gaussian elimination can be performed on the system (2.13) without interchanges; consequently A has an LU factorization.

Remark 1. As we can easily deduce from the proof of Lemma 1, the hypothesis (2.14) expresses the requirement that all diagonal elements of the matrix coefficients A in (2.13) to be non-zero, which guarantees the existence of LU decomposition.

3 Stability conditions

To establish conditions of stability for the linear difference equations expressed by (2.13), we will use in our analysis the Lax-Richtmyer definition of

stability, expressed in terms of norm $\|\cdot\|_\infty$ (see [7, Chapter 5] and references therein). Equation (2.13) may be rewritten in a more convenient form as

$$\varphi^i = A^{-1}B\varphi^{i-1} + A^{-1}d^i \quad i = 2, 3, ..., M \tag{3.1}$$

(the existence of A^{-1} will be proved in the proof of Proposition 1 below). Moreover, the matrix A can be written in the form

$$A = D(I + D^{-1}G) \tag{3.2}$$

where $D = diag(a_1, c_1 + 2c_2 - \frac{1}{2\xi}, \cdots, c_1 + 2c_2 - \frac{1}{2\xi}, a_1)$ and $G = A - D$. Thus, noting $a_2 = c_1 + 2c_2 - \frac{1}{2\xi}$, we have

$$\mathrm{D}^{-1}G = \begin{pmatrix} 0 & -\frac{1+c_2+c_3}{a_1} & 0 & \cdots & 0 & 0 & 0 \\ -\frac{c_2}{a_2} & 0 & -\frac{c_2}{a_2} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\frac{c_2}{a_2} & 0 & -\frac{c_2}{a_2} \\ 0 & 0 & 0 & \cdots & 0 & -\frac{1+c_2+c_3}{a_1} & 0 \end{pmatrix}.$$

The sum of each line in matrix $D^{-1}G$ is written in the vector v below (recall that $a_1 = c_1 + 2c_2 + c_3 + c_0 + 1 - \frac{1}{2\xi}$ and $a_2 = c_1 + 2c_2 - \frac{1}{2\xi}$)

$$v = \left[-\frac{1 + c_2 + c_3}{a_1}, -2\frac{c_2}{a_2}, \cdots, -2\frac{c_2}{a_2}, -\frac{1 + c_2 + c_3}{a_1}\right]. \tag{3.3}$$

Let's denote by

$$v_{max} = \max\{|-(1 + c_2 + c_3)|, |-2c_2|\} \quad \text{and} \quad v_{min} = \min\{|a_1|, |a_2|\}.$$

Now we are able to prove the following result with respect to the stability in matrix equation (3.1).

Proposition 1. *Suppose that* $v_{min} - v_{max} > 0$. *If*

$$\frac{\alpha\xi}{v_{min} - v_{max}} < \frac{\varepsilon}{2} \tag{3.4}$$

then the equation (3.1) *is stable. Otherwise, it is unstable.*

Proof. The proof is reduced to check the inequality $\|A^{-1}B\|_\infty < 1$. We begin by determining an estimate for $\|D^{-1}G\|_\infty = \max|v|$, wherefrom we easily derive the estimate

$$\|D^{-1}G\|_\infty < \frac{v_{max}}{v_{min}}. \tag{3.5}$$

The estimate (3.5) allows now to prove the existence of A^{-1}. Indeed, since by hypothesis we have assumed that $v_{max} < v_{min}$ than $\|D^{-1}G\|_\infty < 1$ which

guarantees that there exist $(I + D^{-1}G)^{-1}$. Moreover, there exist A^{-1} and $A^{-1} = (I+D^{-1}G)^{-1}D^{-1}$. Using the well known inequality: $\|(I+D^{-1}G)^{-1}\|_\infty < \frac{1}{1-\|D^{-1}G\|_\infty}$ and making use of (3.2), it follows that

$$\|A^{-1}\|_\infty \leq \|(I + D^{-1}G)^{-1}\|_\infty \|D^{-1}\|_\infty < \frac{1}{1 - \|D^{-1}G\|_\infty}\|D^{-1}\|_\infty. \quad (3.6)$$

How the inequality $\|D^{-1}G\|_\infty < 1$ imply that $1-\|D^{-1}G\|_\infty > 1-\frac{v_{max}}{v_{min}} > 0$, we easily deduce now that

$$0 < \frac{1}{1 - \|D^{-1}G\|_\infty} < \frac{v_{min}}{v_{min} - v_{max}}.$$

Since $\|D^{-1}\|_\infty \leq \frac{1}{v_{min}}$ and involving the above estimate, from (3.6) we finally obtain

$$\|A^{-1}\|_\infty < \frac{1}{v_{min} - v_{max}}. \quad (3.7)$$

Now we turn our attention to matrix B. Analyzing the matrix B lines, it follows that

$$\|B\|_\infty = \max\{2c_1, c_1\} = 2\frac{\alpha\xi}{\varepsilon}. \quad (3.8)$$

Summing up and making use of (3.7)-(3.8) we derive the following estimate

$$\|A^{-1}B\|_\infty \leq \|A^{-1}\|_\infty \|B\|_\infty < \frac{1}{v_{min} - v_{max}}\|B\|_\infty,$$

which, owing to (3.4), leads us to the estimate $\|A^{-1}B\|_\infty < 1$ as we claimed at beginning of proof.

Remark 2. The hypothesis $v_{min} > v_{max}$ in Proposition 1 derives from the necessity to have a strict sub-unitary estimation for $\max|v|$ (see relation (3.3)). A large part of numerical experiments presented in the next section are designed to support this theoretical aspect.

4 Numerical experiments

The aim of this Section is to present numerical experiments implementing the conceptual algorithm **Alg_1-IMBDF_dbc**. Corresponding to input data T, b, M, N, we have used several different values, while, for the model's parameters, we have started with the values: $\xi = .5$, $\alpha = 1.0e+1$ and $c_0 = 1.0e-3$.

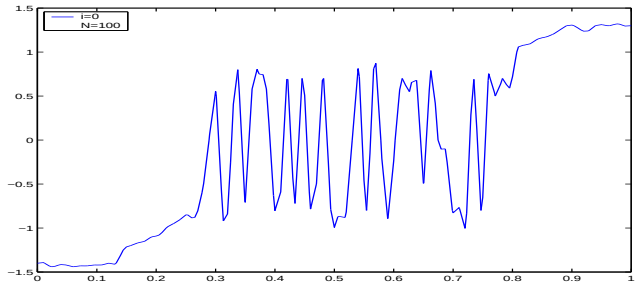

Figure 1: *The initial conditions* φ_0

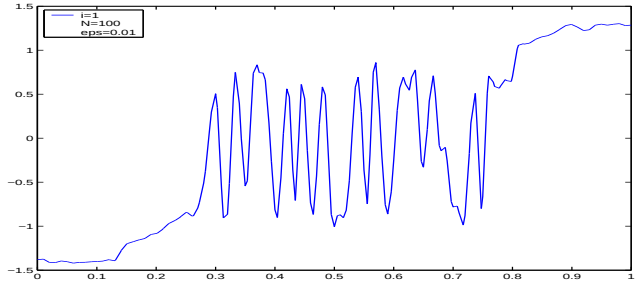

Figure 2: *The approximate solution* $z(\varepsilon,\cdot)$ *of the Cauchy problem* (1.5)

The initial values $\varphi_0(x_j),\ j=1,2,...,N$, plotted in Figure 1, were computed via Matlab function `csapi(fi0)` - cubic spline interpolant, corresponding to the following input data:

fi0=[-1.4 -1.4 -1.44 -1.42 -1.42 -1.44 -1.43 -1.43 -1.42 -1.42 -1.4 -1.4 -1.25 ...
-1.2 -1.17 -1.15 -1.1 -1.08 -1.0 -.95 -.9 -.85 -.88 -.6 .0 .5 -.92 -.25 .8 -.7 ...
.58 .75 .58 -.63 -.59 .69 -.72 .7 -.59 -.5 .7 -.79 -.87 -.88 .0 .72 -.8 .81 ...
.0 -.89 .0 .7 .55 .68 -.49 .79 .0 -.1 -.8 -.78 -.83 .69 .8 .68 .5 .7 .59 1. ...
1.08 1.1 1.15 1.17 1.2 1.25 1.3 1.3 1.25 1.24 1.3 1.31 1.3 1.32 1.3 1.3];

Now (see (2.5)) we are able to calculate the vector $\left(z\left(\varepsilon,\varphi_0(x_j)\right)\right)_{j=\overline{1,N}}$, plotted in Figure 2, and the vector $\varphi^1=\left(\varphi_j^1\right)_{j=\overline{1,N}}$ (see $(1.4)_3$).

We will present now some numerical experiments regarding the *stability* of the matrix equation (3.1), established by Proposition 1. For the first tests, we have set: $T=1$, $b=2$, $M=100$, $N=100$ and the values at boundary given by: $w(t_i,0)=-725$, $w(t_i,b)=0$, $i=1,2,...,M$. We can verify that $v_{min}-v_{max}=1.69e+3>0$ and $\frac{\alpha\xi}{v_{min}-v_{max}}-\frac{\varepsilon}{2}=-0.0021<0$. Consequently, all hypothesis in Proposition 1 are satisfied and then we are

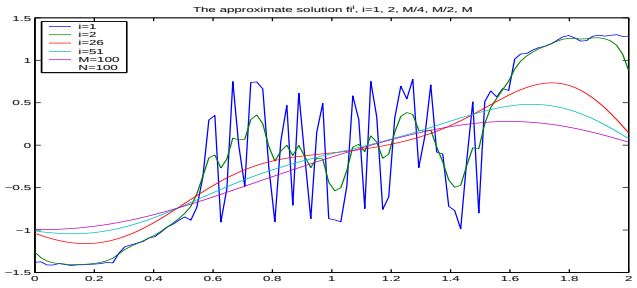

Figure 3: *Example of numerical stability:* φ^i *at different levels of time*

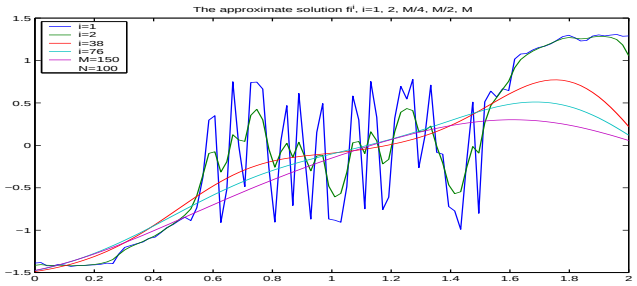

Figure 4: *Example of numerical stability:* φ^i *at different levels of time*

in a stability case. The shape of the graphs plotted in Figures 3 shows that it really is. Changing in the above settings only the value of M, again we are in a stable case. Analyzing the graph in Figure 4 we find a slight improvement of stability in the boundary point $x_1 = 0$.

Taking now $T = 2$, $b = 2$, $M = 50$, $N = 50$ and $\alpha = 1.0e + 2$, one can check that $\frac{\alpha\xi}{v_{min} - v_{max}} - \frac{\varepsilon}{2} = 0.00224 > 0$ which means that the hypothesis (3.4) in not verified, i.e., the numerical scheme (3.1) is unstable (see Figure 5). Changing $\xi = 0.75$ and $c_0 = 1.0e+3$, we get $\frac{\alpha\xi}{v_{min} - v_{max}} - \frac{\varepsilon}{2} = 0.0054 > 0$. So, again we are in a unstable case (see Figure 6). Let's remark that the instability of the solution occurred after a slight change for α, ξ and c_0. This highlights the strong dependence of approximation scheme regarding physical parameters.

We turn to numerical stability conditions changing $w(t_i, 0) = 72.5$ and $w(t_i, b) = -72.5$, $i = 1, 2, ..., M$. We get again a stable case and the numerical results, obtained by algorithms **Alg_1-IMBDF_dbc** were plotted in Figure 7. Analyzing the approximations near 0 and b, we observe a good stability which makes it suitable to be used in the numerical analysis of the

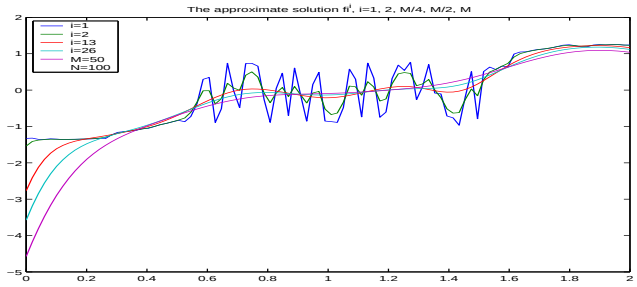

Figure 5: *Example of a numerical instability*

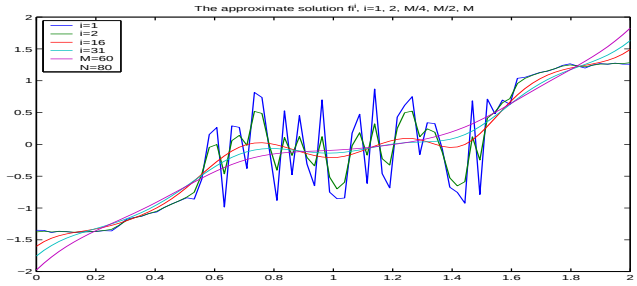

Figure 6: *Example of a slight numerical instability*

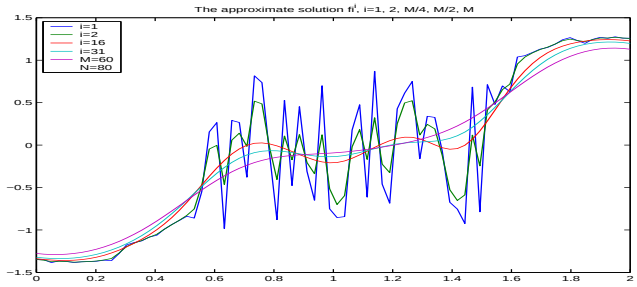

Figure 7: *Example of numerical stability: φ^i at different levels of time*

boundary optimal control problems governed by (1.1).

5 Conclusions

Analyzing the numerical results in terms of physical phenomena (see figures 3-7), we find that *the phase function* distribution say that the instability of the portion of material will disappear. Moreover, the numerical experiments depicted in figure 7, for example, highlight the theoretical meaning assigned to the unknown function φ and the *zone of separation* between material phases.

The numerical solution obtained by this way can be considered as an admissible one for the appropriate boundary optimal control problem (from this perspective, compare figures 4, 5 and 7 in terms of stability). Generally, the numerical method considered here can be used to approximate the solution of a nonlinear parabolic phase-field system containing a general nonlinear part. Not the least, let's remark that conditions of stability are sustained by both theory and numerical experiment and that are significantly dependent on the physical parameters.

References

[1] S. Allen, J.W. Cahn. A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening. *Acta Metall.* 27:1084-1095, 1979.

[2] V. Arnăutu, C. Moroşanu. Numerical approximation for the phase-field transition system. *Intern. J. Com. Math.* 62:209-221, 1996.

[3] T. Benincasa, C. Moroşanu. Fractional steps scheme to approximate the phase-field transition system with non-homogeneous Cauchy-Neumann boundary conditions. *Numer. Funct. Anal. & Optimiz.* 30:199-213, 2009.

[4] G. Caginalp, X. Chen. Convergence of the phase field model to its sharp interface limits. *Euro. Jnl of Applied Mathematics* 9:417-445, 1998.

[5] O. Cârjă, A. Miranville, C. Moroşanu. On the existence, uniqueness and regularity of solutions to the phase-field system with a general regular potential and a general class of nonlinear and nonhomogeneous boundary conditions. *Nonlinear Anal.*, 113:190-208, http://dx.doi.org/10.1016/j.na.2014.10.003, 2015.

[6] Gh. Iorga, C. Moroşanu, I. Tofan. Numerical simulation of the thickness accretions in the secondary cooling zone of a continuous casting machine. *Metalurgia International* XIV:72-75, 2009.

[7] C. Moroşanu. *Analysis and optimal control of phase-field transition system: Fractional steps methods.*, Bentham Science Publishers, http://dx.doi.org/10.2174/9781608053506112010l, 2012.

[8] C. Moroşanu, Ana-Maria Moşneagu. On the numerical approximation of the phase-field system with non-homogeneous Cauchy-Neumann boundary conditions. Case 1D. *ROMAI J.* 9:91-110, 2013.

[9] C. Moroşanu. Numerical analysis of an iterative scheme of fractional steps type associated to the nonlinear phase-field equation in Caginalp's model endowed with non-homogeneous dynamic boundary conditions. *The 22-nd Conference on Applied and Industrial Mathematics, CAIM 2014, Bacău, Romania, September 18-21*, 2014.