FGS 2024

French-German-Spanish Conference on Optimization

June 18-21, 2024

Universidad de Oviedo. Gijón (Asturias)

Proceedings

www.unioviedo.es/fgs2024

Sponsors

Departamento de Matemáticas
Universidad de Oviedo

EPI Gijón
Universidad de Oviedo

visita gijón Convention Bureau

accenture

Fundación Cajastur

Universidad de Oviedo

SëMA
Sociedad Española
de Matemática Aplicada

AMPL

FGS 2024

French-German-Spanish Conference on Optimization

Universidad de Oviedo

Gijón
June 18-21, 2024

Gallego R.; Mateos M. (coords). (2024). Libro de Resúmenes del FGS 2024 (French-German-Spanish Conference on Optimization). Universidad de Oviedo.

Servicio de Publicaciones de la Universidad de Oviedo
Edificio de Servicios - Campus de Humanidades
ISNI: 0000 0004 8513 7929
33011 Oviedo - Asturias
985 10 95 03 / 985 10 59 56
servipub@uniovi.es
www.publicaciones.uniovi.es

I.S.B.N.: 978-84-10135-30-7

FOREWORD

The first edition of the of the French-German Congress of Optimization started back in 1980 in Oberwolfach (Germany) as a meeting of French and German mathematicians who shared a common interest in different aspects of mathematical optimization. Since that edition, researchers of other countries have participated in this congress and, starting in 1998 a third country is invited to participate in the organization.

On this occasion, we have had the honour to be chosen to host it in Spain, so this is a French-German-Spanish conference (FGS2024, https://www.unioviedo.es/fgs2024/). It took place in the Campus of Gijón of the Universidad de Oviedo, in the Spanish region of Asturias.

In this proceedings book you can find the extended abstracts of some of the communications presented at the conference. They include a variety of topics related to optimization, ranging from classical ones --such as optimal control, inverse problems or mathematical programming-- to the latest insights into machine learning and artificial intelligence provided by the application to control theory to the study of neural odes, as generalizations of neural networks.

We would like to thank all the people that have made this congress possible: all the authors and coauthors of presentations, the scientific committee, the technical secretariat of the Foundation Universidad de Oviedo, the students that have volunteered, and our sponsors.

The local organizing committee from the Universidad de Oviedo

Scientific Committee

- Anne Auger, *INRIA Saclay - Île-de-France*
- Eduardo Casas, *Universidad de Cantabria*
- Antonin Chambolle, *Université Paris – Dauphine*
- Elena Fernández, *Universidad de Cádiz*
- Enrique Fernández-Cara, *Universidad de Sevilla*
- Didier Henrion, *LAAS-CNRS Toulouse*
- Marco Antonio López-Cerdá, *Universidad de Alicante*
- Alexander Martin, *Friedrich-Alexander-Universität Erlangen-Nürnberg*
- Helmut Maurer, *University of Münster*
- Sabine Pickenhain, *Brandenburgische Technische Universität*
- Jérôme Renault, *Toulouse School of Economics*
- Dolores Romero-Morales, *Copenhagen Business School*
- Michael Ulbrich, *Technische Universtät München*
- Andrea Walter, *Humboldt-Universität zu Berlin*
- Hasnaa Zidani, *INSA-Rouen*

Local Organizing Committee

- Mariano José Mateos Alberdi, *Universidad de Oviedo*
- Pedro Alonso Velázquez, *Universidad de Oviedo*
- Rafael Gallego Amez, *Universidad de Oviedo*
- Jorge Jiménez Meana, *Universidad de Oviedo*
- Antonio Palacio Muñiz, *Universidad de Oviedo*
- Set Pérez González, *Universidad de Oviedo*
- Maria Luisa Serrano Ortega, *Universidad de Oviedo*
- Jesús Suárez Pérez del Río, *Universidad de Oviedo*

Technical Secretariat

- Adriana Suárez Paredes, *Fundación Universidad de Oviedo*
- Vanessa Cuenco Díaz, *Fundación Universidad de Oviedo*

Contents

French-German-Spanish Conference on Optimization
Gijón, June 18-21, 2024
(pp. 11-16)

Controllability of neural ODEs for data classification

Antonio Álvarez-López
Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Madrid, Spain.

Abstract

In this work, we explore the capacity of neural ordinary differential equations (ODEs) for supervised learning from a control perspective. Specifically, we rely on the property of simultaneous controllability and explicitly construct the controls that achieve this as piecewise constant functions in time.

First, we analyze the expressivity of the model for cluster-based classification by estimating the number of neurons required for the classification of a set constituted by N points. We consider a worst-case scenario where these points are independently sampled from $U([0,1]^d)$. Assuming only that the initial points are in general position, we propose an algorithm that classifies clusters of d points simultaneously, employing $O(N/d)$ neurons.

Secondly, we examine the impact of the architecture, determined by the depth p and width L, for interpolating a set of N pairs of points. Our findings reveal a balance where L scales as $O(1+N/p)$. For the autonomous model, with constant controls ($L=0$), we relax the problem to approximate controllability of N pairs of points, establishing an explicit error decay with respect to p. Finally, we extend the problem to the approximate control of measures in the Wasserstein space, finding another balance between p and L.

1. Introduction

Supervised learning is one of the main paradigms in machine learning. Given some spaces $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}^m$ with $d, m \geq 1$, the problem can be formulated as the approximation of an unknown function $f : \mathcal{X} \to \mathcal{Y}$ using a parametric model built from the information contained in a training dataset $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N \subset \mathcal{X} \times \mathcal{Y}$, where $\mathbf{y}_n = f(\mathbf{x}_n)$ for all n.

Neural networks constitute a widely used class of models, and among them, residual networks have been shown to be particularly effective. A residual neural network, defined for a fixed depth $L \in \mathbb{N}$, operates as a discrete system given by:

$$\mathbf{x}_{l+1} = \mathbf{x}_l + \sum_{i=1}^{p} \mathbf{w}_{l,i}\sigma(\mathbf{a}_{l,i} \cdot \mathbf{x}_l + b_{l,i}), \qquad l = 0, \dots, L, \tag{1.1}$$

where $\mathbf{x}_l \in \mathbb{R}^d$ is the sequence of states, $\cdot$ denotes the scalar product, and:

- $\mathbf{w}_{l,i}, \mathbf{a}_{l,i} \in \mathbb{R}^d$ and $b_{l,i} \in \mathbb{R}$ are the parameters;
- p is the width of the model;
- σ is a predefined nonlinearity, frequently the Rectified Linear Unit (ReLU) function, defined by:

$$\sigma(z) = \max\{z, 0\}, \qquad \text{for } z \in \mathbb{R}. \tag{1.2}$$

Neural ODEs are essentially the continuous-time limit of residual networks [5]. They are obtained by multiplying the nonlinear term in (1.1) by a constant $h > 0$ and taking the limit when $h \to 0$, resulting in:

$$\dot{\mathbf{x}} = \sum_{i=1}^{p} \mathbf{w}_i(t)\, \sigma(\mathbf{a}_i(t) \cdot \mathbf{x} + b_i(t)), \qquad t \in (0, T), \tag{1.3}$$

where the parameters can now be seen as p control functions $(\mathbf{w}_i, \mathbf{a}_i, b_i)_{i=1}^p \subset L^\infty\left((0,T), \mathbb{R}^{2d+1}\right)$, for some $T > 0$. Note that the time horizon T does not play a major role, since equation (1.3) admits a time-rescaling property: one can equivalently fix $T = 1$ and absorb a factor T into $\mathbf{w}_i$.

One of the main advantages of neural ODEs is that they enable the reinterpretation and study of various machine learning paradigms using the tools from differential equations and dynamical systems [10]. For instance, data classification can be formulated as a problem of simultaneous control of the system (1.3). The

 ISBN 978-84-10135-30-7

objective is to design p controls that drive every initial data point $\{\mathbf{x}_n\}_{n=1}^N \subset \mathbb{R}^d$ to its corresponding target point via the flow map at time T of the system (1.3).

To facilitate the geometric interpretation of the dynamics, achieve a layered structure similar to (1.1), and reduce the problem to finite dimensions, it is often assumed that the controls are piecewise constant in time [7, 9]. The discrete network's depth can then be interpreted as the number of distinct values that these controls take, and each of the finite-jump discontinuities, whose total number we denote by L, corresponds to a layer transition.

Within each layer $t \in (t_{k-1}, t_k) \subset (0, T)$, the controls $\mathbf{a}_i(t) \equiv \mathbf{a}_i \in \mathbb{R}^d$ and $b_i(t) \equiv b_i \in \mathbb{R}$ define p hyperplanes $H_1, \dots, H_p$. The ReLU function in (1.2) then activates or deactivates the corresponding half-spaces:

$$H_i^+ := \{\mathbf{x} \in \mathbb{R}^d : \mathbf{a}_i \cdot \mathbf{x} + b_i > 0\} \quad \text{and} \quad H_i^- := \mathbb{R}^d \setminus H_i^+, \qquad \text{for all } i = 1, \dots, p, \tag{1.4}$$

Meanwhile, each control $\mathbf{w}_i(t) \equiv \mathbf{w}_i \in \mathbb{R}^d$ determines a vector field acting solely on the points inside the half-space H_i^+. The total field in (1.3) acts on each point $\mathbf{x} \in \mathbb{R}^d$ as a weighted superposition of the form $\sum_{i=1}^p \operatorname{dist}(\mathbf{x}, H_i^-)\mathbf{w}_i$, where the i-th term is null when $\mathbf{x} \in H_i^-$. By appropriately defining the controls, we can thus fix any hyperplane H_i in $\mathbb{R}^d$ and generate three basic dynamics, as represented in Figure 1.

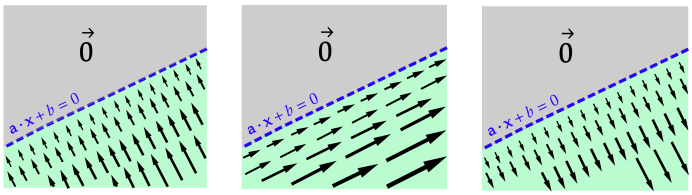


Fig. 1 Basic movements that we can generate: Compression, laminar motion, expansion (from left to right).

2. Controlled cluster-based classification

First, we address binary classification, where $\mathcal{Y} = \{1, 0\}$. In this context, the values y_n are commonly referred to as labels. We associate the two labels with a pair of target regions that are linearly separable and form a partition of $\mathbb{R}^d$. For example, the two half-spaces defined by $x^{(k)} \neq 1$. Our goal is to design controls for the neural ODE that generate a flow mapping each initial point $\mathbf{x}_n$ to the corresponding target region $x^{(k)} > 1$ or $x^{(k)} < 1$.

Furthermore, for optimal classification, the complexity of the model, represented by the number of neurons defining the network, should not grow excessively large. By fixing $p = 1$ in (1.3), the complexity is thus determined solely by the number of discontinuities in the controls over time:

$$\dot{\mathbf{x}} = \mathbf{w}(t)\, \sigma(\mathbf{a}(t) \cdot \mathbf{x} + b(t)), \qquad t \in (0, T). \tag{2.1}$$

In [7], classification of any finite dataset was achieved through a constructive algorithm that leverages the nonlinear dynamics of (2.1) to simultaneously control the N points inductively. The main result in this work is the following:

Theorem 2.1 *Let $N \geq 1$, $d \geq 2$, and $T > 0$. Consider any dataset $\{(\mathbf{x}_n, y_n)\}_{n=1}^N \subset \mathbb{R}^d \times \{1, 0\}$ with $\mathbf{x}_n \neq \mathbf{x}_m$ if $n \neq m$. Then, there exists a piecewise constant control $(\mathbf{w}, \mathbf{a}, b) \in L^\infty\big((0, T), \mathbb{R}^{2d+1}\big)$ such that the flow map Φ_T generated by (2.1) satisfies, for all $n = 1, \dots, N$:*

$$\Phi_T(\mathbf{x}_n)^{(1)} > 1 \quad \text{if } y_n = 1, \qquad \text{and} \qquad \Phi_T(\mathbf{x}_n)^{(1)} < 1 \quad \text{if } y_n = 0,$$

Furthermore, the number of discontinuities in the controls is $L = 3N$.

Theorem 2.1 opens new pathways for methodologies in data classification. However, it requires high complexity since the number of neurons scales with N due to the inductive nature of the algorithm. In [1], we propose new algorithms that consider the spatial structure of the data distribution to reduce the number of parameters needed. Specifically, by assuming that the points are randomly sampled from $U([0, 1]^d)$—a worst-case scenario of pure noise—we construct controls that provide the following probabilistic bound on the model's depth:

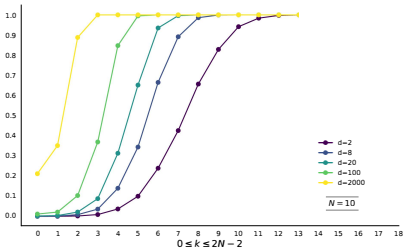
Theorem 2.2 *Let $N \geq 1$, $d \geq 2$, and $T > 0$. Consider any dataset $\{(\mathbf{x}_n, y_n)\}_{n=1}^{2N}$ with $\mathbf{x}_n \sim U([0, 1]^d)$ and $y_n \in \{1, 0\}$ for all n, satisfying $\#\{n : y_n = 1\} = \#\{n : y_n = 0\} = N$. Then, there exist a direction $j \in \{1, \dots, d\}$, a*

piecewise constant control $(\mathbf{w}, b) \in L^\infty\left((0,T), \mathbb{R}^{d+1}\right)$ *and* $\mathbf{a} \in \{\mathbf{e}_1, \dots, \mathbf{e}_d\}$*, such that the flow map* Φ_T *generated by* (2.1) *satisfies, for all* $n = 1, \dots, 2N$*:*

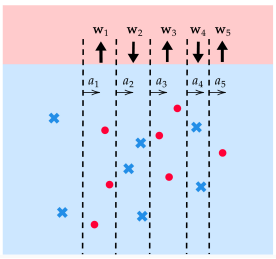
$$\Phi_T(\mathbf{x}_n)^{(j)} < 1 \quad \text{if } y_n = 1, \quad \text{and} \quad \Phi_T(\mathbf{x}_n)^{(j)} > 1 \quad \text{if } y_n = 0.$$

Furthermore, the number of discontinuities L follows the probability distribution, for $0 \le k \le 2N-2$*,*

$$\mathbb{P}(L \ge k) = \left(\sum_{p=\lceil \frac{k+1}{2} \rceil}^{N} \binom{N-1}{p-1}^2 + \sum_{p=\lceil \frac{k}{2} \rceil}^{N-1} \binom{N-1}{p} \binom{N-1}{p-1} \right)^d 2^d \binom{2N}{N}^{-d}. \tag{2.2}$$



(a) Visualization of (2.2) for $N = 10$ and different values of d.



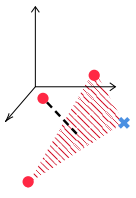
(b) Representation of the algorithm for classification from Theorem 2.2.

Fig. 2

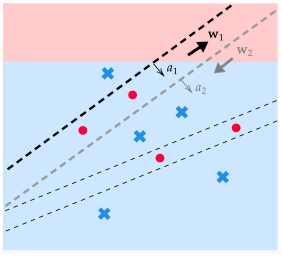
The maximum number of $L = 2N - 2$ discontinuities corresponds to the configuration where the $2N - 1$ points lie on a single line and are interspersed according to their labels. Although these scenarios are typically unrealistic, they hold a positive probability in Theorem 2.2 due to the strong constraint on $\mathbf{a}$. However, if we assume that the points are in general position, meaning no $d + 1$ points lie on the same hyperplane (see figure 3a), we can build new controls that refine the maximum value of L:

Theorem 2.3 *Let* $d \ge 2$*,* $N \ge 1$*, and* $T > 0$*. Consider any dataset* $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N \subset \mathbb{R}^d \times \{1, 0\}$ *in general position and any direction* $j \in \{1, \dots, d\}$*. Then, there exists a piecewise constant control* $(\mathbf{w}, \mathbf{a}, b) \in L^\infty\left((0,T), \mathbb{R}^{2d+1}\right)$ *with* $L = 4\lceil m/d \rceil - 1$ *discontinuities, where* $m = \min\left(\#\{n : y_n = 1\}, \#\{i : y_n = 0\}\right)$*, such that the flow map generated by* (2.1) *satisfies, for all* $n = 1, \dots, N$*:*

$$\Phi_T(\mathbf{x}_n)^{(j)} < 1 \quad \text{if } y_n = 1 \quad \text{and} \quad \Phi_T(\mathbf{x}_n)^{(j)} > 1 \quad \text{if } y_n = 0.$$



(a) General position setting.



(b) Representation of the algorithm for classification from Theorem 2.3.

Fig. 3

3. Interplay between depth and width

3.1. In simultaneous control

As an extension of Theorem 2.1, the property of simultaneous control was also proven in [7] by constructing the necessary controls in (2.1):

Theorem 3.1 *Let $N \geq 1$, $d \geq 2$, and $T > 0$. Consider any dataset $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N \subset \mathbb{R}^d$ with $\mathbf{x}_n \neq \mathbf{x}_m$ and $\mathbf{y}_n \neq \mathbf{y}_m$ for $n \neq m$. Then, there exists a piecewise constant control $(\mathbf{w}, \mathbf{a}, b) \in L^\infty\left((0,T), \mathbb{R}^{2d-1}\right)$ such that the flow map Φ_T generated by (2.1) satisfies:*

$$\Phi_T(\mathbf{x}_n) = \mathbf{y}_n, \qquad \text{for all } n = 1, \dots, N.$$

Furthermore, the number of discontinuities in the controls is $L = 4N$.

In our second work [2], we focus on the role that the architecture can play in this task by allowing the width to be $p \geq 1$ and studying its interplay with the depth L. Our findings reveal a balancing trade-off between these two parameters, as shown in the following result:

Proposition 3.2 *Let $N \geq 1$, $d \geq 2$, and $T > 0$. Consider any dataset $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N \subset \mathbb{R}^d$ with $\mathbf{x}_n \neq \mathbf{x}_m$ and $\mathbf{y}_n \neq \mathbf{y}_m$ for $n \neq m$. Then, for any $p \geq 1$, there exist piecewise constant controls $(\mathbf{w}_i, \mathbf{a}_i, b_i)_{i=1}^p \subset L^\infty\left((0,T), \mathbb{R}^{2d+1}\right)$ such that the flow map Φ_T generated by (1.3) satisfies:*

$$\Phi_T(\mathbf{x}_n) = \mathbf{y}_n, \qquad \text{for all } n = 1, \dots, N.$$

Furthermore, the number of discontinuities in the controls is $L = 2\left\lceil \frac{N}{p} \right\rceil - 1$.

We can see that as the width p increases, the parameter L decreases at the same rate, indicating that both play a similar role in the steering process. However, whenever $p \geq N$, the constructed control will exhibit only one switch ($L = 1$), which precludes a complete transition to the autonomous model

$$\dot{\mathbf{x}} = \sum_{i=1}^{p} \mathbf{w}_i \sigma(\mathbf{a}_i \cdot \mathbf{x} + b_i) \tag{3.1}$$

This is because the proof is algorithmically divided into two phases, represented in Figure 4. First, we control $d-1$ coordinates of each batch of p points, and then we control the remaining coordinate. Therefore, at least one discontinuity is inevitable to transition between these two phases.

Fig. 4 Left: Step 1. Control of $d-1$ coordinates. Right: Step 2. Control of the remaining coordinate.

Motivated by this observation, we now pose the following question:

Is it possible to achieve exact control using $L = 0$ discontinuities?

There are some remarks that can be made as a first approach:

1. **Semi-autonomous neural ODE**: If we consider the semi-autonomous neural ODE where only the controls b_i depend on time,

$$\dot{\mathbf{x}} = \sum_{i=1}^{p} \mathbf{w}_i \sigma(\mathbf{a}_i \cdot \mathbf{x} + b_i(t)), \tag{3.2}$$

we can adapt the proof of Theorem 3.2, obtaining the same result and the same number of discontinuities for some controls $(b_i)_{i=1}^p \subset L^\infty((0,T), \mathbb{R})$, but with constant $(\mathbf{w}_i, \mathbf{a}_i)_{i=1}^p \subset \mathbb{R}^{2d}$.

2. **High dimensions**: When $d > N$, the second step in the proof of Theorem 3.2 can be omitted because we can find a new basis of $\mathbb{R}^d$ in which each point $\mathbf{x}_n$ shares the first coordinate with its target $\mathbf{y}_n$. Thus, we reduce L to $2\left\lceil \frac{N}{p} \right\rceil - 2$.

3. **Probabilistic**: Additionally, we can estimate the probability that the points will appear in certain spatial configurations that facilitate their autonomous control. For instance, if $\mathbf{x}_n$ and $\mathbf{y}_n$ are randomly sampled from $U([0,1]^d)$ for all $n = 1, \dots, N$, then with probability P bounded as

$$1 \geq P \geq 1 - \left[1 - \frac{1}{\sqrt{2}} \left(\frac{e}{2N}\right)^N\right]^d \to 1,$$

there exist p controls $(\mathbf{w}_i, \mathbf{a}_i, b_i)_{i=1}^p \subset \mathbb{R}^{2d+1}$ such that $\Phi_T(\mathbf{x}_n) = \mathbf{y}_n$.

In general, another option is to relax the problem statement to approximate controllability, which means allowing a uniform error $\varepsilon > 0$ that can be made arbitrarily small. Thus, we can obtain the following result:

Theorem 3.3 *Let $N \geq 1$, $d \geq 2$, and $T > 0$. Consider any dataset $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N \subset \mathbb{R}^d$ with $\mathbf{x}_n \neq \mathbf{x}_m$ for $n \neq m$. For each $p \geq 1$, there exist controls $(\mathbf{w}_i, \mathbf{a}_i, b_i)_{i=1}^p \subset \mathbb{R}^{2d+1}$ such that the flow map Φ_T generated by* (1.3) *satisfies*

$$\sup_{n=1,\dots,N} \left|\mathbf{y}_n - \Phi_T(\mathbf{x}_n)\right| \leq C \frac{\log_2(\kappa)}{\kappa^{1/d}},$$

where $\kappa = (d+2)dp$ is the number of parameters in the model, and $C > 0$ is a constant independent of p.

The strategy consists of applying an approximation theorem for shallow neural networks in the space of Lipschitz functions with respect to the uniform norm, providing explicit convergence rates, as derived from [3]. The vector field to be approximated will be a time-independent Lipschitz field whose integral curves guide each input point $\mathbf{x}_n$ in $\mathcal{D}$ to its corresponding target $\mathbf{y}_n$ within a fixed time T. The construction of this field is described in Figure 5.

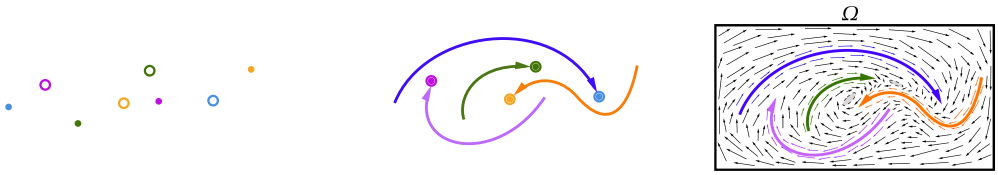


Fig. 5 Construction of a Lipschitz field which interpolates $\mathcal{D}$ in a compact domain Ω that contains all the points and curves.

3.2. In neural transport

As an extension of the results we present in this section, we also consider the reformulation of the model (1.3) as a semilinear hyperbolic equation, known as the neural transport equation:

$$\partial_t \mu + \operatorname{div}_x(\mathbf{V}(\mathbf{x})\mu) = 0, \qquad \text{with} \quad \mathbf{V}(\mathbf{x}) = \sum_{i=1}^p \mathbf{w}_i \sigma(\mathbf{a}_i \cdot \mathbf{x} + b_i). \tag{3.3}$$

This equation defines the evolution of a measure μ in $\mathbb{R}^d$ following an advection vector field $\mathbf{V}$ given by the neural ODE. The case of N initial data points is recovered by taking N Dirac deltas as the base measure, which evolve according to the characteristic equation given by (1.3).

We will work in the space $\mathcal{P}_{ac}^c(\mathbb{R}^d)$ of compactly supported, absolutely continuous probability measures in $\mathbb{R}^d$, with the metric given by the Wasserstein distance, which is rooted in the theory of optimal transport. For any pair of measures $\mu, \nu \in \mathcal{P}_{ac}^c(\mathbb{R}^d)$ and $q \geq 1$, the Wasserstein-q distance between μ and ν is defined by

$$\mathcal{W}_q(\mu, \nu) := \left(\min_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} |\mathbf{x} - \mathbf{y}|^q \, d\gamma(\mathbf{x}, \mathbf{y})\right)^{1/q}, \tag{3.4}$$

where $\Pi(\mu, \nu)$ is the space of all couplings of μ and ν:

$$\Pi(\mu, \nu) := \left\{\gamma \in \mathcal{P}_{ac}^c(\mathbb{R}^d \times \mathbb{R}^d) \mid \gamma(\cdot \times \mathbb{R}^d) = \mu(\cdot) \quad \text{and} \quad \gamma(\mathbb{R}^d \times \cdot) = \nu(\cdot)\right\}.$$

Since the vector field $\mathbf{V}(\mathbf{x})$ in (3.3) is Lipschitz in $\mathbf{x}$, the classic Cauchy-Lipschitz theorem guarantees that the curve $\mu(t)(\cdot) \coloneqq \Phi_t(\cdot;\theta)\#\mu_0$ in $\mathcal{P}_{ac}^c(\mathbb{R}^d)$ is well-defined, where $\Phi_T\#\mu_0$ denotes the pushforward measure under Φ_T.

The objective now is to study the controllability problem of the equation (3.3), aimed at transforming one given probability measure into another, up to an arbitrarily small error ε. As in simultaneous control, the case with $p = 1$ was resolved for both the total variation metric in [8] and the Wasserstein-1 space in [7]. In the latter work, the following result was obtained:

Theorem 3.4 *Let $d \geq 2$ and $T > 0$. For any $\mu_0, \mu_* \in \mathcal{P}_{ac}^c(\mathbb{R}^d)$ and $\varepsilon > 0$, there exists a piecewise constant control $(\mathbf{w}, \mathbf{a}, b) \in L^\infty\left((0,T), \mathbb{R}^{2d+1}\right)$ such that the solution $\mu(t)$ of (3.3), taking μ_0 as initial condition, satisfies*

$$\mathcal{W}_1(\mu(T), \mu_*) < \varepsilon.$$

In our work [2], we study the case with $p \geq 1$ for the uniform measure in the hypercube $[0,1]^d$ as the target. The control algorithm we develop is explicit and allows us to obtain an explicit expression for the number of discontinuities L in terms of p, d, and the order of Wasserstein q:

Theorem 3.5 *Let $d \geq 2$ and $T > 0$. For any $\mu_0 \in \mathcal{P}_{ac}^c(\mathbb{R}^d)$, $\varepsilon > 0$, $q \in \left[1, \frac{d}{d-1}\right)$, and $p \geq 1$, there exist piecewise constant controls $(\mathbf{w}_i, \mathbf{a}_i, b_i)_{i=1}^p \subset L^\infty\left((0,T), \mathbb{R}^{2d+1}\right)$ such that the solution $\mu(t)$ of (3.3), taking μ_0 as the initial condition, satisfies*

$$\mathcal{W}_q(\mu(T), \mu_*) < \varepsilon,$$

and the number of discontinuities in the controls is

$$L = \left\lceil \frac{2d}{p} \right\rceil + \left\lceil \frac{1}{p-d+1} \left(\frac{3^{1+d/q}\sqrt{d}}{\varepsilon} \right)^{\frac{d}{1+d/q-d}} \right\rceil - 1.$$

As a final remark, when $q = 1$ then the number of discontinuities simplifies to:

$$L = \left\lceil \frac{2d}{p} \right\rceil + \left\lceil \frac{1}{p-d+1} \left(\frac{3^{1+d}\sqrt{d}}{\varepsilon} \right)^{d} \right\rceil - 1.$$

Acknowledgements

The author A. Álvarez-López has been funded by contract FPU21/05673 from the Spanish Ministry of Universities and supported by the Government of Madrid under the multiannual agreement with UAM for the Excellence of University Research Staff within the context of the V PRICIT Programme.

References

[1] Antonio Álvarez-López, Rafael Orive-Illera, and Enrique Zuazua. Optimized classification with neural ODEs via separability. *Preprint arXiv:2312.13807*, 2023.

[2] Antonio Álvarez-López, Arselane Hadj Slimane, and Enrique Zuazua. Interplay between depth and width for interpolation in neural ODEs. *Preprint arXiv:2401.09902*, 2024.

[3] Francis Bach. Breaking the Curse of Dimensionality with Convex Neural Networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

[4] Borjan Geshkovski and Enrique Zuazua. Turnpike in optimal control of PDEs, ResNets, and beyond. *Acta Numerica*, 31:135–263, 2022. Cambridge University Press.

[5] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural Ordinary Differential Equations. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6572–6583, Curran Associates Inc., 2018.

[6] Domènec Ruiz-Balet, Elisa Affili, and Enrique Zuazua. Interpolation and approximation via Momentum ResNets and Neural ODEs. *Systems & Control Letters*, 162:105182, 2022.

[7] Domènec Ruiz-Balet and Enrique Zuazua. Neural ODE Control for Classification, Approximation, and Transport. *SIAM Review*, 65(3):735–773, 2023.

[8] Domènec Ruiz-Balet and Enrique Zuazua. Control of neural transport for normalising flows. *J. Math. Pures Appl. (9)*, 181:58-90, 2024.

[9] Stefano Massaroli, Michael Poli, Jinkyoo Park, Atsushi Yamashita, and Hajime Asama. Dissecting neural ODEs. *Advances in Neural Information Processing Systems*, 33:3952–3963, 2020.

[10] E Weinan. A Proposal on Machine Learning via Dynamical Systems. *Communications in Mathematics and Statistics*, 5:1-11, 2017.

French-German-Spanish Conference on Optimization
Gijón, June 18-21, 2024
(pp. 17-22)

The study of an inverse problem for a fluid-solid interaction model in one-dimension

Jone Apraiz[1], **Anna Doubova**[2], **Enrique Fernández-Cara**[3], **Masahiro Yamamoto**[4]

1. *jone.apraiz@ehu.eus Universidad del País Vasco, Spain*
2. *doubova@us.es Universidad de Sevilla, Spain*
3. *cara@us.es Universidad de Sevilla, Spain*
4. *myama@next.odn.ne.jp The University of Tokyo, Japan*

Abstract

This article is a sample of what has been done in the research work [1] and has been presented in the French-German-Spanish Conference on Optimizaion 2024, that has been held in Gijón, Spain. In the work [1] we considered a one-dimensional fluid-solid interaction model governed by the Burgers equation with a time varying interface. There, we studied the inverse problem of determining the shape of the interface from Dirichlet and Neumann data at one end point of the spatial interval.

In this article we display the main results we have obtained in [1] in order to establish uniqueness property and some conditional stability estimates. We also show a brief outline of the proofs, where we have used and adapted some lateral estimates that rely on appropriate Carleman and interpolation inequalities.

1. Introduction

We will consider a nonlinear system that models the interaction of a one-dimensional fluid evolving in $(-1,1)$ and a solid particle. It will be assumed that the velocity of the fluid is governed by the viscous Burgers equation at both sides of the point mass location $y = p(t)$. For simplicity, it will be accepted that the fluid density is constant and equal to 1 and the solid particle has unit mass.

For any p at least in $C^0([0,T])$ satisfying $|p(t)| \leq 1$ for all $t \in [0,T]$, let us introduce the open sets

$$Q(p) = \{(x,t) \in \mathbb{R}^2 : -1 < x < 1,\ x \neq p(t),\ 0 < t < T\},$$
$$Q_\ell(p) = \{(x,t) \in Q(p) : x < p(t)\} \text{ and } Q_r(p) = \{(x,t) \in Q(p) : x > p(t)\}.$$

On the other hand, the *jump* of the function f at the point x will be denoted in the sequel by $[f](x)$, that is,

$$[f](x) := \lim_{s \to 0^+} f(x+s) - \lim_{s \to 0^-} f(x+s).$$

We will consider fluid-particle systems of the form

$$\begin{cases} w_t - w_{xx} + ww_x = 0, & (x,t) \in Q(p), \\ w(p(t),t) = p'(t), \quad [w_x](p(t),t) = p''(t), & t \in (0,T), \\ w(-1,t) = \alpha(t), \quad w(1,t) = \eta(t), & t \in (0,T), \\ w(x,0) = w_0(x), & x \in (-1,1), \\ p(0) = q_0, \quad p'(0) = q_1, \end{cases} \tag{1.1}$$

where (at least) $w_0 \in L^2(-1,1)$, $\alpha, \eta \in C^0([0,T])$, $|q_0| < 1$ and $q_1 \in \mathbb{R}$.

Here, $w(x,t)$ is the velocity of the fluid particle located at x at time t, $p(t)$ is the position occupied by the particle at time t and α and η are Dirichlet data. It is assumed that w_0, q_0 and q_1 are initial data respectively for the fluid velocity, the particle position and the particle velocity.

The first condition at $x = p(t)$ in (1.1) means that the velocity of the fluid and the solid mass coincide at this point. In the second condition, we state *Newton's law:* the force exerted by the fluid on the particle equals the product of the particle mass and its acceleration. Thus, if we introduce the notation $u := w|_{Q_\ell(p)}$ and $v := w|_{Q_r(p)}$, the jump condition at the points $(p(t),t)$ can be written in the form

$$(v_x - u_x)(p(t),t) = p''(t), \quad t \in (0,T). \tag{1.2}$$

The previous system can be viewed as a preliminary simplified version of other more complicate and more realistic models in higher dimensions that we plan to analyze in the future. For example, it is meaningful

 ISBN 978-84-10135-30-7

to consider a system governed by the Navier-Stokes equations around a moving sphere that interacts with the fluid.

As far as we know, the first works where the simplified model (1.1) has been considered are [7] and [6]. There, the authors allowed the spatial variable to take any value in $\mathbb{R}$ instead of $(-1, 1)$. In particular, in [7], the authors proved the existence and uniqueness of a solution and described its large-time behavior for just one solid mass submerged in the fluid. In [6], similar result were established in the case of various rigid bodies immersed in the fluid. These results were later extended to a multi-dimensional framework in [5]. Let us also mention that the controllability properties of a system similar to (1.1) have been analyzed in [3] and [4].

First, we will identify the so called *direct problem* for (1.1):

Direct problem - Given the data $T > 0$, $w_0 \in H^1(-1,1)$, $q_0 \in (-1,1)$, $q_1 \in \mathbb{R}$, $\alpha \in C^0([0,T])$ and $\eta \in C^0([0,T])$, find the solution (w,p) to (1.1).

It can be shown that, if $\alpha(0) = w_0(-1)$, $\eta(0) = w_0(1)$ and $|q_0| + |q_1| + \|w_0\|_{H^1(-1,1)}$ is sufficiently small, there exists a solution (w,p) to (1.1) with $w \in C^0([0,T]; H^1(-1,1))$ and $p \in H^2(0,T)$; see for example [4, Theorem 1.1]. In fact, the result in [4] only states that $p \in C^1([0,T])$. However, the regularity of the restrictions of w to $Q_\ell(p)$ and $Q_r(p)$ shows that the a.e. defined function $t \mapsto [w_x](p(t),t)$ is square-integrable and, consequently, $p'' \in L^2(0,T)$.

The *inverse problem* related to system (1.1) we are interested in is the following:

Inverse problem - Given the data $T > 0$, $q_0 \in (-1,1)$, $q_1 \in \mathbb{R}$ and $\alpha \in C^0([0,T])$ and the observation β with $\beta(t) = w_x(-1,t)$ for $t \in (0,T)$, find $\eta := w(1,\cdot)$.

In what follows, we will frequently use solutions (w,p) with

$$p \in H^2(0,T), \;\; u := w|_{Q_\ell(p)} \in H^2(Q_\ell(p)) \;\text{ and }\; v := w|_{Q_r(p)} \in H^2(Q_r(p)).$$

In the work [1] we have studied related uniqueness and stability properties. In particular, we will show the answers we have obtained to questions like the following:

Global uniqueness - Let (w_i, p_i) be a solution to (1.1) associated to some T, q_0, q_1 and α for $i = 1, 2$. Assume that the corresponding observations coincide at $x = -1$, that is, $w_{1,x}(-1,t) = w_{2,x}(-1,t)$ for $0 < T_1 < t < T_2 < T$. Then, do we have $p_1 = p_2$ and $w_1 = w_2$?

Global stability - Let (w_i, p_i) be as before and set $\beta_i := w_{i,x}(-1,\cdot)$ and $\eta_i = w_i(1,\cdot)$ for $i = 1, 2$. Is there any estimate of the kind

$$\|\eta_1 - \eta_2\|_{L^\infty(T_1,T_2)} + \|p_1 - p_2\|_{L^\infty(T_1,T_2)} \leq \phi(\|\beta_1 - \beta_2\|_{L^\infty(0,T)})$$

for some continuous function $\phi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ satisfying $\lim_{s\to 0^+} \phi(s) = 0$?

In order to show the main results we have obtained, in Section 2 we will see a preliminary fundamental lemma that plays a key role in the proof of conditional stability. It provides estimates of the traces on the interface $x = p(t)$ of the difference of two solutions to (1.1) in terms of the boundary data and observations. Then, in Section 3, we will show the stability estimate and then the uniqueness of the lateral inverse problem corresponding to the system satisfied in the left part $Q_\ell(p)$ of the whole domain. By reflection, similar results have been fulfilled by the solution to the system satisfied in $Q_r(p)$. On the other hand, section 4 contains a global stability and uniqueness result for the inverse problem in the whole domain $Q(p)$. Lastly, in section 5 we will see some possible open problems we could study in a future work.

2. Preliminaries

In this section we see a preliminar lemma that is crucial for the proof of a local stability property that will be established in Section 3 (see Proposition 3.1).

Lemma 2.1 *Let us assume that*

$$\begin{cases} u_t - u_{xx} + au_x + bu = 0, & (x,t) \in Q_\ell(p), \\ u(-1,t) = \alpha(t), \quad u_x(-1,t) = \beta(t), & t \in (0,T), \end{cases} \tag{2.1}$$

with $a, b \in L^\infty(Q_\ell(p))$, $\alpha \in H^{3/2}(0,T)$, $\beta \in H^{1/2}(0,T)$, $u \in H^2(Q_\ell(p))$ *and there exist constants* $M > 0$ *and* $\delta \in (0,1)$ *such that*

$$\|u\|_{H^2(Q_\ell(p))} \leq M, \ \ \|p\|_{H^2(0,T)} \leq M \ \text{ and } \ |p(t)| \leq 1 - \delta \quad \forall\, t \in [0,T]. \tag{2.2}$$

Then:

a) For any $\epsilon > 0$, *there exist constants* $K_\epsilon > 0$ *and* $\theta_\epsilon \in (0,1)$ *such that*

$$|u(p(t),t)| \leq \frac{K_\epsilon}{\left|\log \frac{1}{k}\right|^{\theta_\epsilon}} \quad \forall\, t \in [\epsilon, T], \tag{2.3}$$

provided α, β *and* k *satisfy*

$$0 \leq \|\alpha\|_{L^2(0,T)} + \|\beta\|_{L^2(0,T)} < k < 1. \tag{2.4}$$

b) In particular, if $\alpha \equiv 0$ *and* $\beta \equiv 0$ *in* $(0,T)$, *then* $u \equiv 0$ *in* $Q_\ell(p)$.

Outline of the proof:

a) The main idea in the proof of this result is the adaptation of some arguments from [8] that rely on Carleman estimates.

- **Step 1**: after two changes of variables, $Q_\ell(p)$ is transformed to $(0,2) \times (0,T)$ and

$$(2.3) \quad \Leftrightarrow \quad |u(0,t)| \leq \frac{K_\epsilon}{\left|\log \frac{1}{k}\right|^{\theta_\epsilon}} \quad \forall\, t \in [\epsilon, T].$$

- **Step 2**: prove an intermediate estimate:

$$|u(0,\bar{t})| \leq \frac{K_{0,\epsilon}}{\left|\log \frac{1}{F_\epsilon}\right|^{\theta_0}} \quad \forall\, \bar{t} \in [\epsilon, T], \tag{2.5}$$

where $\theta_0 \in (0,1)$ is independent of ϵ and

$$F_\epsilon := \sup_{x \in [1,2]} \left(\|u(x,\cdot)\|_{L^2(\epsilon,T)} + \|u_x(x,\cdot)\|_{L^2(\epsilon,T)} \right).$$

The tools we have used here are: changes of variables, a cut-off function, global Carleman inequality, optimization and Sobolev interpolation.

- **Step 3**: we find a lateral estimate of F_ϵ, that is, for every $\epsilon > 0$, there exist constants $C_\epsilon > 0$ and $\theta_\epsilon \in (0,1)$ such that

$$G_\sigma \leq F_\epsilon \leq C_\epsilon M^{1-\theta_\epsilon} \left(\|\alpha\|_{L^2(0,T)} + \|\beta\|_{L^2(0,T)} \right)^{\theta_\epsilon} + C_\epsilon \left(\|\alpha\|_{L^2(0,T)} + \|\beta\|_{L^2(0,T)} \right),$$

where $G_\sigma^2 := \sup_{x \in [1,2]} \left(\|\hat{u}(x,\cdot)\|^2_{L^2(\sigma, T^2/(2\bar{t}))} + \|\hat{u}_x(x,\cdot)\|^2_{L^2(\sigma, T^2/(2\bar{t}))} \right)$ after doing the change of variables $\hat{t} = \dfrac{T}{2\bar{t}}\, t$ for $\bar{t} \in [\epsilon, T]$ and $\hat{u}(x,\hat{t}) = u(x,t)$ for $(x,t) \in (0,2) \times (0,T)$. Then, we use an interpolation inequality when $0 \leq \|\alpha\|_{L^2(0,T)} + \|\beta\|_{L^2(0,T)} < k < 1$.

b) We have used (2.3) with $\alpha = 0$, $\beta = 0$ and k arbitrarily small.

□

3. Lateral estimates and uniqueness

This section is devoted to show the stability and uniqueness results of (1.1) we obtained on the left part of the domain, $Q_\ell(p)$. Later, we extend these results to $Q_r(p)$ and obtain similar results in the whole domain $Q(p)$.

Assume that

$$\begin{cases} u_t^i - u_{xx}^i + u^i u_x^i = 0, & (x,t) \in Q_\ell(p_i), \\ u^i(-1,t) = \alpha^i(t), \quad u_x^i(-1,t) = \beta^i(t), & t \in (0,T), \\ u^i(p_i(t),t) = p_i'(t), & t \in (0,T), \end{cases} \tag{3.1}$$

for $i = 1, 2$. Let us formulate an inverse problem concerning the left part of the domain:

Lateral uniqueness in $Q_\ell(p)$: Let (u^i, p_i), $i = 1, 2$ be two solutions to (3.1) in $Q_\ell(p_i)$. Assume that the corresponding observations coincide at the boundary $x = -1$, that is,

$$u_x^1(-1,t) = u_x^2(-1,t) \text{ in some time interval } (T_1, T_2).$$

Then, do we have $p_1 = p_2$ in $(0,T)$ and $u^1 = u^2$ in $Q_\ell(p)$ with $p = p_1 = p_2$?

The following proposition may be viewed as a first conditional stability result:

Proposition 3.1 (Local stability for the lateral inverse problem) *Let us assume that*

$$\|u^i\|_{H^2(Q_\ell(p_i))} \le M, \ \ \|p_i\|_{H^2(0,T)} \le M \ \text{ and } \ |p_i(t)| \le 1 - \delta \ \ \forall t \in [0,T]$$

for some $\delta \in (0,1)$. Also, let us assume that $0 < \epsilon < \bar{t} < T$ and

$$0 \le D := \|\alpha^1 - \alpha^2\|_{L^2(0,T)} + \|\beta^1 - \beta^2\|_{L^2(0,T)} < k < 1.$$

Then there exist R_ϵ, $R_0 > 0$ and $\mu_\epsilon \in (0,1)$ such that

$$\|p_1 - p_2\|_{L^\infty(\epsilon,T)} \le \frac{R_\epsilon}{\left(\log \frac{1}{k}\right)^{\mu_\epsilon}} + R_0 |p_1(\bar{t}) - p_2(\bar{t})|. \tag{3.2}$$

Outline of the proof:

- First, we assume $p_1(t) \le p_2(t)$ for $t \in (t_0, t_1) \subset [\epsilon, T]$ and set $h := p_1 - p_2$.
- For all $t \in (t_0, t_1)$ one has
$$h'(t) \le \frac{K_\epsilon}{\left(\log \frac{1}{D}\right)^{\theta_\epsilon}} + 2Mh(t).$$
- We apply the previous Lemma 2.1 to $u^1 - u^2$ in combination with the Mean Value Theorem for $u^2(\cdot, t)$:
$$\frac{1}{2}\frac{d}{dt}|h(t)|^2 \le C\,|h(t)|^2 + \frac{K_\epsilon}{\left(\log \frac{1}{D}\right)^{2\theta_\epsilon}}.$$
- Using Gronwall's Lemma:
$$|h(t)|^2 \le \frac{K_\epsilon}{\left(\log \frac{1}{D}\right)^{2\theta_\epsilon}} + C\,|h(\bar{t})|^2 \quad \forall\, t \in [\epsilon, T],$$
and this implies (3.2).

□

Remark 3.2 Let the assumptions in Proposition 3.1 be satisfied. Also, suppose that

$$\|u^i\|_{W^{2,\infty}(Q_\ell(p_i))} \le M.$$

Then, it can be ensured that for every $\epsilon > 0$ there exist K_ϵ, K_0 and $\theta_\epsilon \in (0,1)$ such that

$$\|p_1' - p_2'\|_{L^\infty(\epsilon,T)} + \|p_1'' - p_2''\|_{L^\infty(\epsilon,T)} \le \frac{K_\epsilon}{\left(\log \frac{1}{D}\right)^{\theta_\epsilon}} + K_0\,|p_1(\bar{t}) - p_2(\bar{t})|. \tag{3.3}$$

□

Corollary 3.3 *Under the assumptions in Proposition 3.1, if* $0 < \bar{t} < T$ *and* $\alpha^1 \equiv \alpha^2$ *and* $\beta^1 \equiv \beta^2$ *in* $(0,T)$, *there exists a constant* $R_0 > 0$ *such that*

$$\|p_1 - p_2\|_{L^\infty(0,T)} \leq R_0 \, |p_1(\bar{t}) - p_2(\bar{t})|, \tag{3.4}$$

where R_0 *is independent of* $\bar{t}$.

Proof:

We can argue as in the proof of Proposition 3.1. Thus, for every $\epsilon > 0$ and every small $k > 0$, we obtain

$$\|p_1 - p_2\|_{L^\infty(\epsilon,T)} \leq \frac{R_\epsilon}{\left(\log \frac{1}{k}\right)^{\mu_\epsilon}} + R_0 |p_1(\bar{t}) - p_2(\bar{t})|.$$

Then, taking $k \to 0$, we see that

$$\|p_1 - p_2\|_{L^\infty(\epsilon,T)} \leq R_0 |p_1(\bar{t}) - p_2(\bar{t})|.$$

Finally, taking $\epsilon \to 0$, we arrive at (3.4). □

Corollary 3.4 (Lateral uniqueness) *In addition to the assumptions in Corollary 3.3, let us assume that* $p_1(\bar{t}) = p_2(\bar{t})$ *for some* $\bar{t} \in (0,T)$. *Then,*

$$p_1 \equiv p_2 \quad in \quad (0,T) \quad and \quad u^1 \equiv u^2 \quad in \quad Q_\ell(p).$$

4. Global estimates and uniqueness

In this section we present global stability and uniqueness results for the inverse problem formulated in Section 1 in the whole domain $Q(p)$.

Theorem 4.1 (Conditional stability) *Let* (w_1, p^1) *and* (w_2, p^2) *be the solutions to* (1.1) *respectively corresponding to the data* u_0, α, m, q_0, q_1, β^i *and* η^i *and set* $\beta^i(t) = w_x^i(-1,t)$ *and* $\eta^i(t) = w_i(1,t)$ *for* $i = 1,2$ *and all* $t \in (0,T)$. *Assume that there exist constants* $\delta, \kappa \in (0,1)$ *and* $M > 0$ *such that* $|p_i(t)| \leq 1 - \delta$ *for all* $t \in (0,T)$,

$$\|u^i\|_{W^{2,\infty}(Q_\ell(p_i))} \leq M, \ \ \|v^i\|_{W^{2,\infty}(Q_r(p_i))} \leq M \ \ (i = 1,2) \ \ and \ \ 0 \leq \|\beta^1 - \beta^2\|_{L^2(0,T)} < \kappa < 1.$$

Then, for every $\epsilon > 0$, *there exist constants* $C_0, C_\epsilon > 0$ *and* $\theta_\epsilon \in (0,1)$ *such that*

$$\|\eta^1 - \eta^2\|_{L^\infty(\epsilon,T)} \leq \frac{C_\epsilon}{|1 + \log|\log \kappa||^{\theta_\epsilon}} + C_0 |\eta^1(\bar{t}) - \eta^2(\bar{t})|, \tag{4.1}$$

for all $\bar{t} \in [\epsilon, T)$.

Main idea of the proof:

- We find estimates of $\|\eta^1 - \eta^2\|_{L^\infty(\epsilon,T)}$ in terms of $p_1 - p_2$, $p_1' - p_2'$ and $p_1'' - p_2''$ and use previous results, mainly:
  - Two changes of variables.
  - Proposition 3.1 of the local stability for the lateral inverse problem.
  - The inequality of Remark 3.2, (3.3).

□

From this result we deduce that, as $\|\beta_1 - \beta_2\|_{L^2(0,T)} \to 0$, the corresponding $\|\eta^1 - \eta^2\|_{L^\infty(\epsilon,T)}$ goes to zero at a logarithmic rate. In particular, we have:

Corollary 4.2 (Global uniqueness) *Let the assumptions in Theorem 4.1 be satisfied and let us assume that* $\beta^1 = \beta^2$ *in* $(0,T)$. *Then*

$$\eta^1 = \eta^2 \quad in \quad (0,T). \tag{4.2}$$

Outline of the proof:

- Given an arbitrary $\epsilon > 0$, take $\bar{t}_\epsilon = 2\epsilon$, and for every $\kappa > 0$, we use previous Theorem 4.1 to obtain:

$$\|\eta^1 - \eta^2\|_{L^\infty(\epsilon,T)} \leq \frac{C_\epsilon}{|1 + \log|\log\kappa||^{\theta_\epsilon}} + C_0|\eta^1(\bar{t}_\epsilon) - \eta^2(\bar{t}_\epsilon)|.$$

- Take $\kappa \to 0$ and $\epsilon \to 0$ in order to get $\|\eta^1 - \eta^2\|_{L^\infty(0,T)} \leq C_0|\eta^1(0) - \eta^2(0)|$.
- Since $\eta^1(0) = \eta^2(0) = w_0(1)$, we have $\eta^1 = \eta^2$ in $(0,T)$.

□

5. Open Problems

Among many problems that can be studied related to different aspects of this work, here we highlight some of them:

1. **Reconstruction of the unknown data:** knowing the observation $\beta(t) = w_x(-1,t)$ for $t \in (0,T)$, we could try to reconstruct the unknown data $\eta := w(1,\cdot)$. Here, we can analyze and compare other works done in reconstruction and see if we can apply similar techniques for the system (1.1).

2. **Take $\epsilon = 0$:** we can try to find a more involved argument in order to take $\epsilon = 0$ in inequality (2.3) of the preliminar Lemma 2.1, and therefore in Proposition 3.1 and Theorem 4.1. We forsee that in this case the stability rate is expected to be weaker than single logarithmic.

3. **Similar model in higher dimensions:** we can try to extend these kind of results to higher dimensions. We can model the problem with the Navier-Stokes equations and set up the corresponding IP, and then, try to obtain some stability and uniqueness results. See for example the model presented in [2] and the inverse problem studied there.

Acknowledgements

The first author was supported by the Grant PID2021-126813NB-I00 funded by MICIU/AEI/10.13039/5011000110 33 and by "ERDF A way of making Europe" and by the grant IT1615-22 funded the Basque Government. The second and third authors were partially supported by grant PID2020-114976GB-I00 funded by MICIU/AEI/10.13039/50110
0011033 (Spain). The fourth author was supported by Grant-in-Aid for Scientific Research (A) 20H00117 and Grant-in-Aid for Challenging Research (Pioneering) 21K18142 of the Japan Society for the Promotion of Science

References

[1] J. Apraiz, A. Doubova, E. Fernández-Cara, M. Yamamoto. *Inverse problems for one-dimensional fluid-solid interaction models*, accepted for publication in Communications on Applied Mathematics and Computation. At this moment it is published in *Arxiv*, 2401.16546, 2024.

[2] A. Doubova, E. Fernández-Cara and J. H. Ortega, *On the identification of a single body immersed in a Navier-Stokes fluid*, Euro. Jnl of Applied Mathematics **18**, (2007), 57–80.

[3] A. Doubova and E. Fernández-Cara, *Some control results for simplified one-dimensional models of fluid-solid interaction*, Mathematical Models & Methods in Applied Sciences **15**, no. 5 (2005), 783–824.

[4] Y. Liu, T. Takahashi, and M. Tucsnak, *Single input controllability of a simplified fluid-structure interaction model*, ESAIM: Control, Optimisation and Calculus of Variations **19** (2013) 20–42.

[5] A. Munnier and E. Zuazua, Large time behavior for a simplified n-dimensional model of fluid-solid interaction, Communications in Partial Differential Equations **30**, no. 3, (2005) 377–417.

[6] J.L. Vázquez and E. Zuazua, Lack of collision in a simplified 1-d model for fuid-solid interaction, Mathematical Models & Methods in Applied Sciences **16**, no. 5, (2006), 637–678.

[7] J.L. Vázquez and E. Zuazua, Large time behavior for a simplified 1D model of fluid-solid interaction, Comm. Partial Differential Equations **28** (2003), no. 9–10, 1705–1738.

[8] M. Yamamoto, Carleman estimates for parabolic equations and applications, Inverse Problems **25** (2009) 123013.

French-German-Spanish Conference on Optimization
Gijón, June 18-21, 2024
(pp. 23-30)

Control and estimation for the design of a smart electrostimulator using Ding et al. model

Toufik Bakir[1], **Bernard Bonnard** [2], **Ilias Boualam** [3], **Mokrane Abdiche** [4]

1. toufik.bakir@u-bourgogne.fr ImVIA, université de Bourgogne, France
2. bernard.bonnard@u-bourgogne.fr Institut de Mathématique de Bourgogne, université de Bourgogne and Mc Tao team, Inria Sophia Antipolis, France
3. mohamedilias.boualam@segula.fr Entreprise Segula Matra Automotive, Trappes, France
4. mokrane.abdiche@segula.fr Entreprise Segula Matra Automotive, Trappes, France

Abstract

Based on the A. V Hill's muscle model (Medicine Nobel prize 1922), mathematical models validated by experiments due to Ding et al. in the 2000's allow to describe the muscular force isometrical contraction due to electrostimulation, taking into account the fatigue. They serve as a model to control and to predict the effect of trains of electrical stimulations, with rest periods aiming to force rehabilitation or reinforcement. In this article we briefly present the main issues of the problem. Two typical training sessions are described related to increase the force or the endurance. Each program is translated into an optimization problem which is analyzed in the sample-data control frame. The parameters of the models split into parameters independent of each individual vs. parameters related mainly to the fatigue, which have to be online estimated. Geometric estimation theory leads to describe a software sensor to make explicit computations. NMPC algorithm vs MPC algorithm can be used to regulate the force.

1. Introduction

Recent mathematical models validate by experiments due mainly to Ding et al. [4–6] and based on the earliest work by A.G. Hill [9] allow to predict the isometric force response to external electrical stimulation, taking into account the fatigue phenomenon due to a long stimulation period. Such models contain two basic nonlinearities which constitute the intricate part of the dynamics. First of all, the ionic conduction and the nonlinear effect of successive pulses on the Ca^{++}concentration. Second, the nonlinearity relating the muscular force response to such concentration, modeled by the Michaelis-Menten-Hill functions, which cause the force saturation called tetany. The control is formed by a sequence of trains of pulses which fit in the frame of sample-data control (digital controls) due to limitation on the interpulse. Our objective is to use the model to construct a smart electrostimulator for force rehabilitation or reinforcement based on two objectives : maximize the force response F_{max} to a single train corresponding exactly to the tetany or an endurance session regulating the force to a reference force e.g. $\frac{F_{max}}{2}$ while minimizing the fatigue. Each training session is limited to 30 minutes since external stimulation causes severe fatigue and even during an endurance session rest periods have to imposed. Besides those objectives they are computational limits related to on-board electronics and cost reduction. In particular integrating the nonlinear dynamics is time consuming and is bypassed by an approximation of the force response.

Hence the first part of this article is to briefly recall an off line formal approximation of the force dynamics to compute F_{max}. It is based on a piecewise linear approximations of the Michaelis-Menten-Hill functions and is fully described in [3]. The second part of this article is to describe an internal input-output model which is used to regulate the force-fatigue to a given level using Model Predictive Control [1, 13] and based on the Ding et al. (nonlinear)model. The final important issue of the project is to estimate the parameters based on preliminary experiments in the industrial realization of the electrostimulator. They are based mainly on the piezoelectric force sensor and the measurements being realized either at the beginning of each training session or during the "rest periods" where the muscle can be in reality stimulated with small intensity and frequency. Geometric control techniques developed in the 90s, see for instance [10], are used to study the observability of the system and to identify the "bad inputs" for which the systems is not observable. In the experiments they correspond roughly to the zero input where no force is produced. Recent geometric estimation techniques allow to identify the parameters of the model. The experiments they can be sorted into two types in the Ding et al. model: four parameters which are not depending on the individuals and four additional parameters related mainly to the fatigue phenomenon and which are depending of each individual and can be time varying. Such parameters can be estimated in the frame of geometric estimation developed in the 2000s and the general techniques presented in [7, 8, 11, 12] based on the construction of normal coordinates and Luenberger-type observers, where the effect of the inputs formed by trains of impulses on the

 ISBN 978-84-10135-30-7

Ca^{++} concentration is described and leads to explicit estimation of the so-called observation space related to such inputs.
Numerical simulations are presented for the MPC algorithm based on the linear parametric model with on-line computation of the parameters and the train of pulses, taking into account the problem of severe fatigue caused by external stimulations.

2. Force-fatigue model

2.1. Ding et al. force-fatigue model

The FES input over a pulse train $[0, T]$ is modelled as a sum of Dirac pulses by

$$t \to \sum_{i=0,n} \eta_i \delta(t - t_i), \tag{2.1}$$

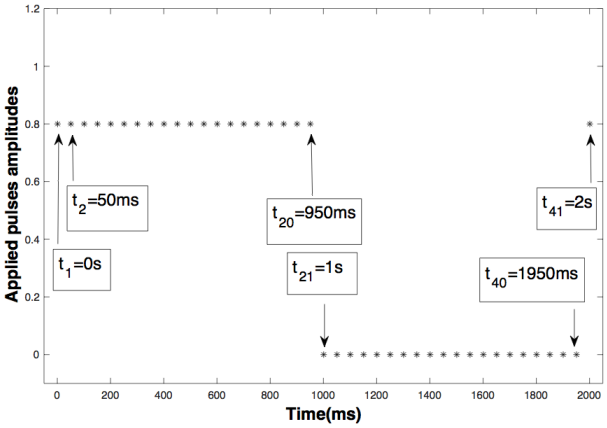


Fig. 1 stimulation period, stimulation and rest sub-periods

where $0 = t_0 < t_1 ... < t_n < t_{n+1} = T$ are the impulses times with $n \in \mathbb{N}$ being fixed (see figure 1 for constant interpulse and stimulation amplitude), η_i being the amplitude of each pulse, which are convexified by taking $\eta_i \in [0, 1]$ and $\delta(. - t_i)$ denoting the Dirac at time t_i. We denote by $I_i = t_i - t_{i-1}$ the interpulse and we have a digital constraint $I_i \geq I_m$ in the problem e.g. $I_m \geq 30ms$ for a train of 10 impulses of around $T = 500ms$. Such a control provides the FES signal taken as the physical input, using a linear filter (first-order linear dynamics).

$$\frac{dE}{dt}(t) + \frac{1}{\tau_c} E(t) = \sum R_i \eta_i \delta(t - t_i) \tag{2.2}$$

so that it takes the form

$$E(t) = \frac{1}{\tau_c} \sum_{i=0,n} R_i e^{-\frac{t-t_i}{\tau_c}} \eta_i H(t - t_i), \tag{2.3}$$

where H is the Heaviside function. $E(t)$ depends upon the time response parameter τ_c and the scaling function R_i depending on parameter $R(0)$ as following:

$$R_0 = 1, R_i = 1 + (R(0) - 1)e^{-(t_i - t_{i-1})/\tau_c}, i = 1, ..., n, \tag{2.4}$$

which codes the memory effect of successive muscle contractions.

The FES signal drives the evolution of the electrical conduction describing the evolution of Ca^{++}-concentration c_N which is related to the force response F. The dynamics being described by

$$\frac{dc_N}{dt}(t) = E(t) - \frac{c_n(t)}{\tau_c}, \tag{2.5}$$

$$\frac{dF}{dt}(t) = -m_2(t)F(t) + m_1(t)A(t) \tag{2.6}$$

where

$$m_1(t) = \frac{c_N(t)}{K_m + c_N(t)}, m_2(t) = \frac{1}{\tau_1 + \tau_2 m_1(t)}. \tag{2.7}$$

Hence six parameters are introduced in the model $(\tau_c, R(0), \tau_1, \tau_2, K_m, A(t))$, where to simplify $(\tau_c, R(0), \tau_1, \tau_2, K_m)$ are fixed parameters and the time variable parameter $A(t)$ is the scaling force parameter which is used to model the fatigue dynamics according to

$$\frac{dA}{dt}(t) = -\frac{A(t) - A_0}{\tau_{fat}} + \alpha_A F(t). \tag{2.8}$$

Tab. 1 Ding et al. model parameters

Symbol	Unit	Value	description
c_N	—	—	Normalized amount of Ca^{2+}-troponin complex
F	N	—	Force generated by muscle
t_i	ms	—	Time of the i^{th} pulse
n	—	—	Total number of the pulses before time t
i	—	—	Stimulation pulse index
τ_c	ms	20	Time constant that commands the rise and the decay of C_N
$R(0)$	—	1.143	Term of the enhancement in C_N from successive stimuli
A	$\frac{N}{ms}$	—	Scaling factor for the force and the shortening velocity of muscle
τ_1	ms	—	Force decline time constant when strongly bound cross-bridges absent
τ_2	ms	124.4	Force decline time constant due to friction between actin and myosin
K_m	—	—	Sensitivity of strongly bound cross-bridges to C_N
A_{rest}	$\frac{N}{ms}$	3.009	Value of the parameter A when muscle is not fatigued
$K_{m,rest}$	—	0.103	Value of the parameter K_m when muscle is not fatigued
$\tau_{1,rest}$	ms	50.95	The value of the parameter τ_1 when muscle is not fatigued
α_A	$\frac{1}{ms^2}$	$-4.0\ 10^{-7}$	Coefficient for the force-model parameter A in the fatigue model
α_{K_m}	$\frac{1}{msN}$	$1.9\ 10^{-8}$	Coefficient for the force-model parameter K_m in the fatigue model
α_{τ_1}	$\frac{1}{N}$	$2.1\ 10^{-5}$	Coefficient for force-model parameter τ_1 in the fatigue model
τ_{fat}	s	127	Time constant controlling the recovery of $(A,\ K_m,\ \tau_1)$

Values of the parameters are reported in the reference [4] (see table 1) in the frame fo Ding et al. experiments, the system formed by (2.5) and (2.6) describing the non-fatigue model, while the additional equation (2.8) is describing the fatigue and depends on two parameters α_A which defines the "slope" of the fatigue evolution while τ_{fat} is the time constant controlling the recovery to the rest point $A_{rest} = A(0)$. The model provides a closed curve $t \to A(t)$ obtained from the fatigue dynamics associated to concatenation of two arcs: the first one associated to the application of the averaged force $F_{averaged} = \frac{1}{T}\int_0^T F(t)dt$ over a pulse train on $[0, T]$ and the recovery arc during the complete rest period where no force is applied.

The main properties of the dynamics of the non-fatigue model is resumed in two lemmas.

Lemma 2.1 *For a pulse train defined by $\sigma = (t_0 = 0, t_1, ..., t_n, t_{n+1} = T, \eta_0, \eta_1, ..., \eta_n)$ the concentration c_N can be written as the superposition of $n + 1$ lobes*

$$c_N(t) = \frac{1}{\tau_c} \sum_{i=0,n} R_i \eta_i (t - t_i) e^{-\frac{t-t_i}{\tau_c}} H(t - t_i) \tag{2.9}$$

which represents a piecewise polynomial-exponential mapping.

Lemma 2.2 *The force dynamics in the non-fatigue case can be written as*

$$\frac{dF}{ds}(s) = c(s) - F(s),$$

using the time reparameterization $ds = m_2(t)dt$ and can be integrated by quadrature using Lagrange formula. This gives an explicit force response $s \to F(s)$ which is smooth with respect to the control parameters and s at each time different of a pulse time t_i.

From which we deduce the following , see [3] for the complete details and numerical simulations.

3. Construction of the approximation of the force response to a single train and the Punch Program in non-fatigue case

3.1. Approximation

According to (2.9) each lobe l_k is given by

$$l_k = R_k \eta_k \frac{t - t_k}{\tau_c} e^{-(t-t_k)/\tau_c} H(t - t_k),$$

the lobe reaches its maximum at $t = t_k + \tau_c$ which is equal to $R_k \eta_k / e$ and is concave on $[t_k, t_k + 2\tau_c]$ and can be approximated by its restriction to $[t_k, t_k + 5\tau_c]$. The restriction of m_1 to one lobe is maximal when the concentration c_N is maximal and we denote t_k^* the corresponding time. Let σ be the sequence defined in (2.9) and assume that the minimal interpulse is such that $I_m \geq \tau_c$.

We divide the the subdivision $t_0 < t_1 < ... < t_n < T$ introducing the intermediate times t_k^* where m_1 and m_2 are respectively approximated by a piecewise linear mapping and a piecewise constant mapping on each subinterval.

This leads to an explicit formula for the force response F on $[0, T]$. Note that this basic partition can be refined to improve the approximation, see [3] for the complete description. The approximation contains the parameters of Ding et al. model.

3.2. Punch program

Using the previous force approximation denoted $F_{approximation}$ one compute a local minimum σ^* over the set of pulse trains σ. The details of the optimization algorithm and the numerical simulations are presented in [3].

4. Endurance session using the force-fatigue model and the MPC algorithm

4.1. Notation 1

The force fatigue model described by (2.5), (2.6), (2.8) is shortly written as

$$\frac{dx}{dt}(t) = X(x(t)) + u(t)Y(x(t)), \tag{4.1}$$

where $x = (c_n, F, A)^T$ is the state variable and $u(t)$ denotes the general input corresponding to the FES signal. Restriction on u are imposed by the physical device: bounds implied by the constraints $\eta_i \in [0, 1]$, sampling times and interpulse constraints. They will be considered as soft constraints in the MPC algorithm and relaxed in the control computations using a quadratic optimization method.

We assume that in the dynamics two variables are observed $y = h(x) = (F, A)$ which defines the construction of the input-output dynamics.

In the MPC algorithm we consider the following discrete linear input-output system

$$\begin{cases} VX(k+1) = MA_k VX(k) + MB_k U(k) \\ y(k) = CX(k) \end{cases} \tag{4.2}$$

where:

$$MA_k = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix}_k, MB_k = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}_k, VX(k) = \begin{pmatrix} F_{m_k} \\ A_{m_k} \end{pmatrix}. \tag{4.3}$$

The parametric model (4.2) results from the identification routine minimizing the criterion:

$$J = \min_{N_i = N_{i_1}, \ldots, N_{i_{max}}} \frac{1}{N_i} \sum_{j=k-N_i}^{k} \left(\begin{pmatrix} F_{mean_j} \\ A_{mean_j} \end{pmatrix} - \begin{pmatrix} F_{m_j} \\ A_{m_j} \end{pmatrix} \right)^2. \tag{4.4}$$

N_i being the backward identification horizon. Figure 2 represents the force, the mean force and the backward identification horizon to be found in order to get the best parametric model. The same figure could be constructed for A. F_{mean_k} and A_{mean_k} are calculated as following:

$$\begin{aligned} F_{mean_k} &= \frac{1}{t_{k+1} - t_k} \int_{t_k}^{t_{k+1}} F(\xi) d\xi \\ A_{mean_k} &= \frac{1}{t_{k+1} - t_k} \int_{t_k}^{t_{k+1}} A(\xi) d\xi \end{aligned} \tag{4.5}$$

The criterion (4.4) traduces the fact that MA and MB are updated at each iteration (MA_k, MB_k). The couple is used in MPC strategy to calculate the control value η_k. The frequency of the stimulation being fixed.

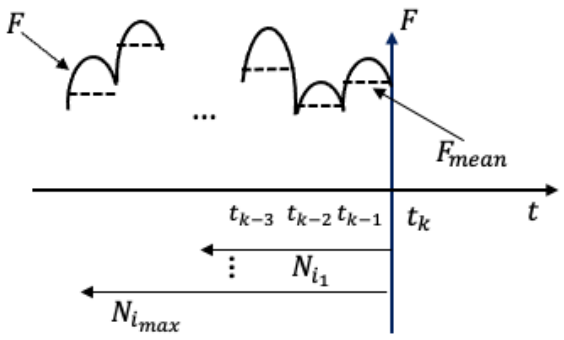


Fig. 2 Force, mean force and identification horizons to identify the parametric model

4.2. MPC Algorithm

We present a version of the algorithm to illustrate the procedure which is classical, and the quadratic cost can be modified. The control constraints have been relaxed and they have to be introduced later to define the true feedback control .

We fix a sequence $k = 1, ..., K$ where N_p is the prediction horizon and an output reference trajectory y_{ref} associated to regulation of the force response F to a fixed level F_{max}/ρ with $\rho > 1$ and a fatigue reference $A_{ref}(.)$. Denoting by $e(k) = (y(k) - y_{ref}(k))$, we minimize a cost of the form

$$J(y,u) = \sum_{k=1,K} \lambda_1 \parallel e(k) \parallel_2^2 + \lambda_2 \parallel \Delta u_k \parallel_2^2, \tag{4.6}$$

where $\Delta u(k)$ is the control increment and λ_i are weighting parameters .

The feedback control $u(k)$ is computed on the horizon K solving the LQ-problem defined by the linear dynamics (4.2) with the quadratic cost (4.6). We implement $u(1)$ and we restart the computations.

The nonlinear system (4.1) is used as a simulation of the data which will be replaced by the experimental data during the endurance session.

5. Estimation of the parameters in the design of the electrostimulator

5.1. Notations and definitions

The force fatigue model is written shortly

$$\frac{dx}{dt}(t) = X(x(t)) + u(t)Y(x(t)), \tag{5.1}$$

where $x = (x_1, x_2, x_3, x_4, x_5)^T = (c_N, F, A, \alpha_A, \tau_{fat})^T$ and u represents the FES input which can be smoothed as $u = u_{smooth}$.

The full system is defined by extending the dynamics with $\frac{d\alpha_A}{dt} = \frac{d\tau_{fat}}{dt} = 0$. We denote by $h = (h_1, h_2) = (F, A)$ the observation mappping.

Fixing a smooth control $u(t)$, the system with $x(0) = 0$ defines a control trajectory pair $(x(.), u(.)$ and we denote in short the Lie derivative $L_{X+uY}h(x(t)) = \frac{d}{dt}h(x(t))$.We denote by $O(x)$ the observation space formed by the iterated functions $\{L^k_{X+uY}h_i; i = 1, 2, k = 0, +\infty\}\}$. The system is called $u-$(weakly) observable if $x \to dO(x)$ is of full rank=dimension of the state space. Given a smooth input u the system is called locally observable if there exists a sequence $0, .., k_1, 0, ...k_2$ so that the mapping $x \to \Phi(x,u) = [h_1(x), ..., L^{k_1}_{X+uY}h_1(x), h_2(x), ..., L^{k_2}_{X+uY}h_2(x)]$ is a diffeomorphism with respect to x for all (x,u) in an nonempty set $\chi \times U$ where U contains the $k-1$ derivatives of u, with $k = max(k_1, k_2)$. We say that $\chi \times U$ is an observable set. The construction of the observer is described in full details in [12] and is presented shortly in the next section.

5.2. Construction of the observer

Assume that the control trajectory pair $(x,u) \in \Omega_x \times \Omega_u \subset \chi \times U$. Perform the nonlinear change of coordinates $z = \Phi(x,u)$ and construct the observer

$$\frac{dz}{dt}(t) = Pro(Az) + \rho(z,u) + S^{-1}K_0(y - Cz). \tag{5.2}$$

The triple $(A, C, \varrho(z,u))$ is obtained writing the system is the coordinates z. The matrix K_0 is chosen so that $(A - K_0C)$ is Hurwitz and $Proj(y,z)$ is the projection operator associated to

$$p(z) = \frac{\| z - z_0 \|^2 - r_\Omega^2}{\alpha^2 + 2\alpha r_\Omega}.$$

The point z_0 is the center of the domain $\Omega_z = B(z_0, r_\Omega)$ contained in $\Phi(\chi \times U)$ and α is a arbitrarily small positive constant. The construction involves block diagonal matrices including S described in [12]. Note that it is introduced in relation with uniform linearization which is related in our construction to the uniform construction of the observable canonical form.

5.3. Geometric application

The experiments show that among the set of parameters the parameters $(\tau_c, R(0))$ are fixed and not depending upon the individual. Hence in particular the Ca^{++} concentration c_N can be taken as the control variable and can be chosen smooth according to a smooth FES-signal or taking the averaged value $c_{Naveraged}(t) = \frac{1}{t}\int_0^t c_N(s)ds$ over any subinterval of the training period. The bad input behavior is related to $c_N = 0$, in computing the inverse mapping of the map $z = \Phi(x,u)$. Hence note that the observer can be turned off imposing that: $\alpha \leq c_N \leq \beta$. At low level of stimulations corresponding to rest period one can rescaled $\alpha \to \in \alpha$ and expand the $F-$dynamics described by the Michaelis-Menten-Hill functions in Taylor Series at $c_N = 0$, at a given order. This will reduce the computational complexity of the Lie derivatives which involve the derivative of the the Michaelis-Menten-Hill functions with respect to c_N and the time derivative of the concentration. A test input function of the form $c_N(t) = a + bsin(\omega t)$ where a, b chosen so that the concentration stays in a arbitrarily band domain.

Additional parameters A_{rest}, K_m are depending upon the individuals and can be estimated using the observer (5.2) during a single train $[0, T]$ using the force sensor only.

6. Simulation results

6.1. System identification

To identify the parametric model which will be used to calculate the MPC based control strategy, we use the Ding et al. model instead of real force and fatigue values (coming from experiments). The parametric model (linear model) will traduce locally the behaviour of the muscle, and needs to be updated for each new interpulse using, in our case, a variable identification moving horizon.

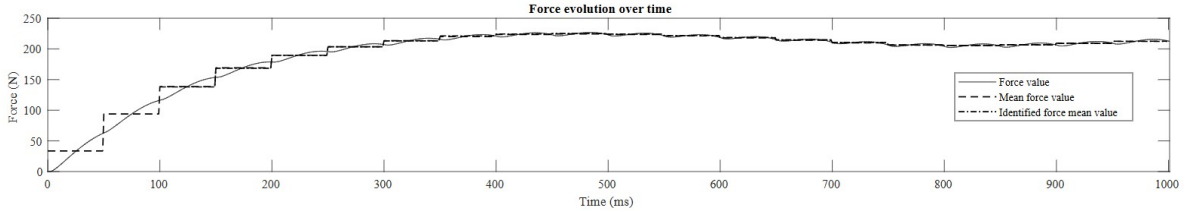


Fig. 3 Evolution of the force (Ding et al. model) over a 1 second stimulation period with a $50ms$ interpulse interval, mean force values (for each interpulse) and identified mean force values over time

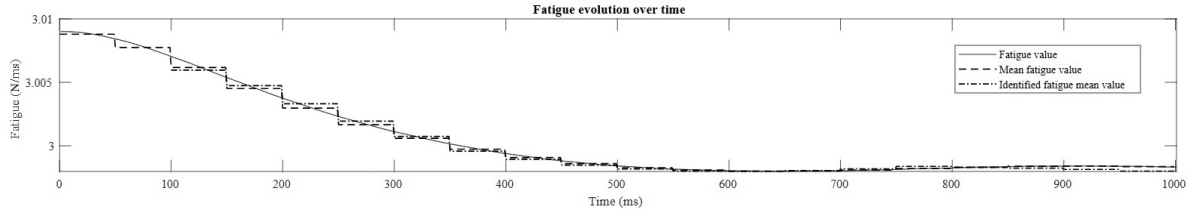


Fig. 4 Evolution of the fatigue (Ding model), mean fatigue values and identified mean fatigue values over time

Figures 3 and 4 show the evolution of the force and the fatigue based on the Ding et al. model over a 1 second stimulation period, with a $50ms$ interpulse interval, respectively. These figures clearly display the lobes generated by this stimulation. The mean values of this force for each interpulse, as well as those obtained by the least squares method, are also shown. The identified mean force value fits well the mean force value over identification horizon.

6.2. Model predictive control

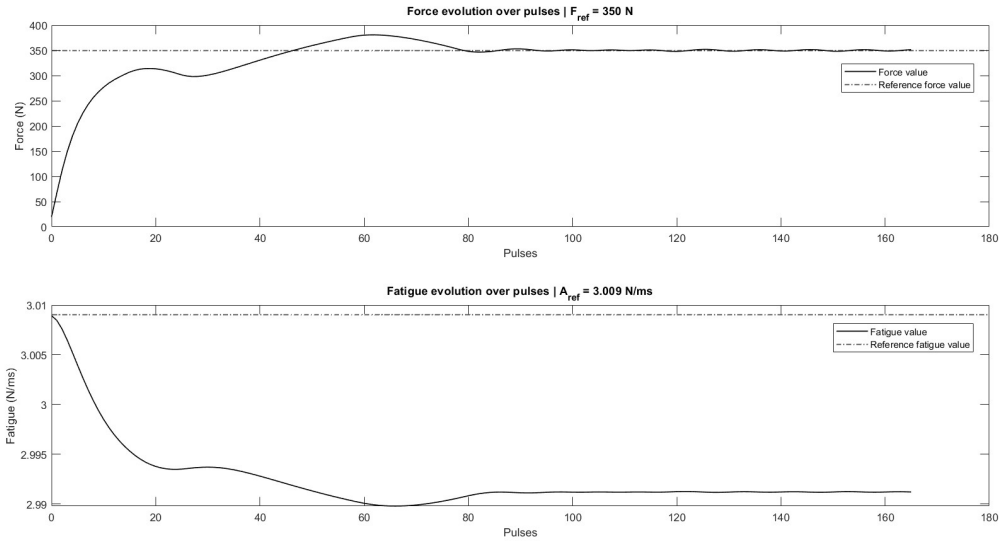


Fig. 5 Evolution of the force and fatigue over pulses at a reference of F=350N

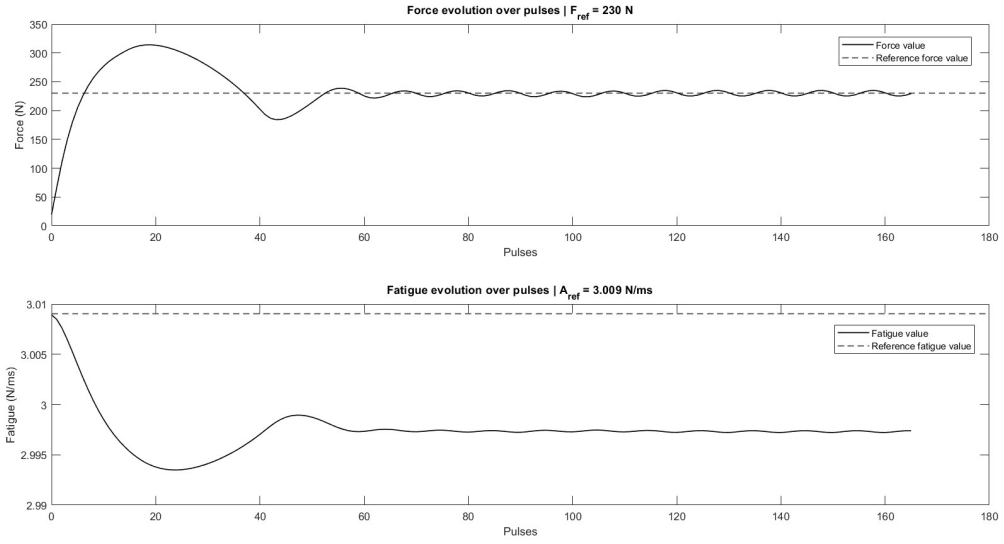


Fig. 6 Evolution of the force and fatigue over pulses at a reference of F=230N

Figures 5 and 6 illustrates the evolution of force and fatigue over pulses, computed from the Ding et al. model, over a stimulation period of 5 seconds with an interpulse interval of 30ms. The reference values for fatigue remain constant at 3.009N/ms, while the reference values for force are respectively 350N (5) and 230N (6). As expected, the MPC strategy allows to fit the force references while minimizing the difference between the fatigue and the fatigue rest value.

7. Conclusion

In this brief article we present the main steps in the design of a smart electrostimulator in relation with the construction of an industrial prototype: model, training sessions and estimation of the parameters using the physical sensors. Numerical simulations are presented for the MPC algorithm implemented to regulate the force and fatigue using a parametric model. The parameters are identified using the data of the Ding et al. model and will be replaced in fine by the experimental data. The Ding et al. model can be used to implement a NMPC algorithm where the parameters are estimated using an observer. But the method is computationally

expensive and MPC algorithm based on linear parametric model can be chosen to tackle computational time while giving good results in terms of force and fatigue control.

References

[1] Toufik Bakir, Bernard Bonnard, Jérémy Rouot. A case study of optimal input-output system with sampled-data control: Ding et al. force and fatigue muscular control model. Netw. Hetero. Media, 14(1) 79–100, 2019.

[2] Toufik Bakir, Bernard Bonnard, Loic Bourdin, Jérémy Rouot. Pontryagin-type conditions for optimal muscular force res-ponse to functional electrical stimulations. J. Optim. Theory Appl., 184 581–602, 2020.

[3] Toufik Bakir, Bernard Bonnard, Sandrine Gayrard, Jérémy Rouot. Finite Dimensional Approximation to Muscular Response in Force-Fatigue Dynamics using Functional Electrical Stimulation Automatica, 2022, ⟨10.1016/j.automatica.2022.110464⟩.

[4] Jun J. Ding, Anthony S. Wexler, Stuart A. Binder-Macleod. Development of a mathematical model that predicts optimal muscle activation patterns by using brief trains. J. Appl. Physiol., 88 917–925, 2000.

[5] Jun J. Ding, Anthony S. Wexler, Stuart A. Binder-Macleod. A predictive fatigue model. I. Predicting the effect of stimulation frequency and pattern on fatigue. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 10(1), 48–58, 2002.

[6] Jun J. Ding, Anthony S. Wexler, Stuart A. Binder-Macleod. A predictive fatigue model. II. Predicting the effect of resting times on fatigue. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 10(1) 59–67, 2002.

[7] Joseph E. Gaudio et al.. Parameter estimation in adaptative control and time-varying systems under a range of excitation conditions, ArXiv:1911.03810v3, November 17th, 2021.

[8] Jean-Paul Gauthier, Ivan Kupka. Deterministic Observation Theory and Applications, Cambridge University Press, 2001.

[9] Rudolf Gesztelyi et al.. The Hill equation and the origin of quantitative pharmacology. Archive for history of exact sciences 66(4) 427–438, 2012.

[10] Alberto Isidori. Nonlinear Control Systems, Springer Verlag London, 1995.

[11] Jean-Baptiste Pomet and Laurent Praly. Adaptative nonlinear regulation: Estimation from the Lyapunov equation, IEEE Trans. on Automatic Control, 37:729-740, October 1992.

[12] Domitilla Del Vecchio and Richard M. Murray. Observability and Local Observer Construction for Unknown Parameters in Linearly and Nonlinearly Parameterized Systems, American Control Conference, 2003.

[13] Yang Wang, Stephen P. Boyd. Fast model predictive control using online optimization, IEEE Transactions on Control Systems Technology, 18 no. 2, 267–278, 2010.

French-German-Spanish Conference on Optimization
Gijón, June 18-21, 2024
(pp. 31-36)

Optimal control for neural ODE in a long time horizon

Jon Asier Bárcena-Petisco
jonasier.barcena@ehu.eus Department of Mathematics, University of the Basque Country UPV/EHU, Spain

Abstract

We study the optimal control, in a long time horizon, of neural ordinary differential equations which are control-affine or whose activation function is homogeneous. When considering the classical regularized empirical risk minimization problem we show that, in long time and under structural assumption on the activation function, the final state of the optimal trajectories has zero training error if the data can be interpolated and if the error can be taken to zero with a cost proportional to the error. These hypotheses are fulfilled in the classification and ensemble controllability problems for some relevant activation and loss functions.

1. Introduction

In this work we study the optimal control of neural ordinary differential equations for a long time horizon. Neural ODE have been used in Machine Learning in the last seven years, a trend started with [6,13]. However, they date back to the 90s, when they were already used for the construction of controls (see the survey [12]) and when their controllability properties were first studied (see, for example, [14] and [11]). The control systems governed by neural ODE have considerably better controllability properties than linear control systems. In fact, as pointed out in [10], for a fixed $d \in \mathbb{N}$, if chosen the right neural ODE we can interpolate an arbitrarily large amount of data in $\mathbb{R}^d$, whereas in linear systems we can at most interpolate an amount of data equal to the dimension of the control. In this paper d denotes the dimension of the space where each element of the dataset is, and N the size of the dataset.

Roughly, the problem under study is the following: given a set of initial values $\mathbf{x} = (x^1, \dots, x^N) \in (\mathbb{R}^d)^N_*$, for:

$$(\mathbb{R}^d)^N_* := \{(x^1, \dots, x^N) \in (\mathbb{R}^d)^N : x^i \neq x^j \ \ \forall i, j \in \{1, \dots, N\} : i \neq j\},$$

we seek to take simultaneously the data set to some target points or regions in $\mathbb{R}^d$ in a given time $T > 0$. This is usually called dataset as it is a set of values. The control problem is important in the context of ensemble controllability. The distance to those targets is measured with an error function (also known as *loss function*). The control is the minimizer of the risk minimization functional, which provides a balance between a small cost for the control and a small value for the loss function at the final state of the optimal trajectory. For a detailed introduction to the notation and its background, I recommend [3, 10].

We study the controllability on control-affine neural networks, which are given by the following equations:

$$\begin{cases} \dot{y}(t) = w(t)\sigma(y(t)) + b(t), \\ y(0) = x, \end{cases} \tag{1.1}$$

for $x \in \mathbb{R}^d$ the initial value, and $\sigma : \mathbb{R}^d \mapsto \mathbb{R}^d$ a Lipschitz function, which is called the *activation function*. The functions (w, b) are the controls and they belong to $L^2(0, T; \mathcal{U})$, for $\mathcal{U}$ defined by:

$$\mathcal{U} := \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times 1}.$$

If we want to emphasize the dependence of (1.1) to the initial value and the control, we write $y(\cdot; x, w, b)$. Similarly, we denote the sequence of solutions of (1.1) for some fixed control (w, b) applied simultaneously to a data set $\mathbf{x}$ as:

$$y(\cdot; \mathbf{x}, w, b) := \left(y(\cdot; x^1, w, b), \dots, y(\cdot; x^N, w, b)\right). \tag{1.2}$$

Since σ is Lipschitz, (1.1) is well-posed by the Cauchy-Lipschitz Theorem.

In addition, we also study more compound neural networks, which are given by the equations:

$$\begin{cases} \dot{y}(t) = r(t)\sigma\left(w(t)y(t) + b(t)\right), \\ y(0) = x. \end{cases} \tag{1.3}$$

 ISBN 978-84-10135-30-7

Here x is the initial value and (r, w, b) is the control, which belongs to $L^2(0, T; \tilde{\mathcal{U}})$, for:

$$\tilde{\mathcal{U}} := X \times \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times 1},$$

for:

$$X \subseteq \{M \in \mathbb{R}^{d\times d} : M_{i,i} \in \{1, -1\},\ \forall i = 1, \dots, d,\ M_{i,j} = 0,\ \forall i \neq j\}. \tag{1.4}$$

In fact, the intensity of the flow is modelled by (w, b), and the direction of the flow, by r. We may take $X = \{I\}$, which makes sense when σ admits negative values. However, we have considered the general setting to have relevant results also for the case in which σ is a positive function; that is, in which $\sigma \geq 0$. We assume that the activation function σ is Lipschitz and homogeneous in the sense that:

$$\sigma(\lambda x) = \lambda\sigma(x),\ \forall \lambda > 0,\ \forall x \in \mathbb{R}^d. \tag{1.5}$$

This includes important *activation functions* such as rectified linear units, which are given by:

$$\sigma(x) = (\max\{x_1, 0\}, \dots, \max\{x_d, 0\}),$$

see [9]; and parametric rectified units, given by:

$$\sigma(x) = (\alpha x_1 1_{x_1<0} + x_1 1_{x_1>0}, \dots, \alpha x_d 1_{x_d<0} + x_d 1_{x_d>0}),$$

see [7]. As in the previous system:

$$y(\cdot; \mathbf{x}, r, w, b) = \big(y(\cdot; x^1, r, w, b), \dots, y(\cdot; x^N, r, w, b)\big), \tag{1.6}$$

where $y(\cdot; x, r, w, b)$ denotes the solutions of (1.3), which is a well-posed system by the Cauchy-Lipschitz Theorem.

For a detailed exposition of the history of this research line, one main consult [1].

2. Main results

2.1. Optimal trajectories for control-affine neural ODE

As stated in the introduction, we study the optimal control of a data set ruled by a neural ODE. To measure how far the data is from the objective we introduce the *error function* (also referred in the literature of Machine Learning as *loss function*) $\mathcal{E} : (\mathbb{R}^d)^N \mapsto \mathbb{R}^+ := [0, \infty)$. We assume that $\mathcal{E}$ is continuous and satisfies the Hypothesis 1, which is later introduced in this section.

This allows to define the *empirical risk minimization functional for a target time T*:

$$J_T(w, b) := \mathcal{E}(y(T; \mathbf{x}, w, b)) + \int_0^T |(w(t), b(t))|^2 dt, \tag{2.1}$$

where y denotes a solution of (1.1) and $|\cdot|$ denotes the Frobenius norm. We denote any minimizer of J_T by (w_T, b_T). Moreover, the trajectories induced by such minimizers, called *optimal trajectories*, are denoted by $y_T(t; \mathbf{x}) := y(t; \mathbf{x}, w_T, b_T)$.

Example 2.1 A usual definition for the error function is:

$$\mathcal{E}(\mathbf{x}) := \frac{1}{N}\sum_{i=1}^{N} E_i(x^i),\ \forall \mathbf{x} \in (\mathbb{R}^d)^N, \tag{2.2}$$

for $E_i(x) = d(x, A_i)$, for d the euclidean distance and for given sets $A_i \subset \mathbb{R}^d$ (that might consist of a single element).

First of all, we recall that the functional J_T has at least a minimizer:

Proposition 2.2 (Existence of minimizers) *Let $\mathcal{E} : (\mathbb{R}^d)^N \mapsto \mathbb{R}^+ := [0, \infty)$ a continuous function, σ a globally Lipschitz continuous function, $T > 0$ and $\mathbf{x} \in (\mathbb{R}^d)^N$. Then, the functional J_T given in* (2.1) *for y given by* (1.2)*, where we consider the solution of* (1.1)*, has at least one minimizer in $L^2(0, T; \mathcal{U})$.*

The main idea of the proof is that J_T is a sum of a positive weakly continuous functional and a positive continuous convex functional.

Let us now present the hypotheses that we consider throughout the paper:

Hypothesis 1 *Let* $\mathbf{x} \in (\mathbb{R}^d)^N_*$, *let* $\mathcal{E} : (\mathbb{R}^d)^N \mapsto \mathbb{R}^+ := [0, \infty)$ *be a continuous function, and let* y *denote* (1.2), *where we consider the solutions of* (1.1). *Then,*

1. *For the data set* $\mathbf{x}$ *there are controls:*

$$(w_*, b_*) \in L^2(0, 1; \mathcal{U}),$$

such that $\mathcal{E}(y(1; \mathbf{x}, w_*, b_*)) = 0$.

2. *There are* $C, \tilde{\varepsilon} > 0$ *both just depending on* $\mathcal{E}$ *such that for all* $\overline{\mathbf{x}} = (\overline{x}^1, \dots, \overline{x}^N) \in (\mathbb{R}^d)^N_*$ *satisfying* $\mathcal{E}(\overline{\mathbf{x}}) < \tilde{\varepsilon}$, *there are some controls* (w, b) *satisfying:*

$$\|(w, b)\|_{L^\infty(0,1;\mathcal{U})} < C\mathcal{E}(\overline{\mathbf{x}}),$$

such that:

$$\mathcal{E}(y(1; \overline{\mathbf{x}}, w, b)) = 0.$$

The first item of Hypothesis 1 is that the error can be taken to 0, a property known in Machine Learning as *interpolation* (see [2]), and the second one is a local controllability of the system.

Remark 2.3 The choice of the target time in Hypothesis 1 is arbitrary. Because of the linearity, if the system is controllable for some time, in this case $T = 1$, it is controllable for any time.

Example 2.4 (Application of Theorem 2.5 to the classification problem) Let us fix $M \in \mathbb{N}$ and consider:

$$\mathbf{x} = \left(x^1, \dots, x^M, x^{M+1}, \dots, x^N\right) \in (\mathbb{R}^d)^N_*,$$

the error function given by (2.2), for:

$$E_i(x) = \begin{cases} (x_1 + 1)1_{x_1 > -1}(x_1), & i = 1, \dots, M, \\ (x_1 - 1)1_{x_1 > 1}(x_1), & i = M + 1, \dots, N, \end{cases}$$

and any neural function σ of the type $\sigma(x) = (\tilde{\sigma}(x_1), \dots, \tilde{\sigma}(x_d))$ such that there is $c > 0$ such that $cs \leq \tilde{\sigma}(s)$ for all $s \geq 0$ and $\tilde{\sigma}(s) \leq cs$ for all $s \leq 0$. The second item of Hypothesis 1 is clearly satisfied, as it suffices to consider $\tilde{\varepsilon} = 1/(2N)$, $b = 0$ and $w(t)x = (2Nc^{-1}\mathcal{E}(\overline{\mathbf{x}})x_1, 0, \dots, 0)$. Thus, Theorem 2.5 implies that if the data can be classified (i.e. if the first item of Hypothesis 1 is satisfied), then by computing the optimal control for a sufficiently large time, the data is sent to the sets $\{x_1 \leq -1\}$ and $\{x_1 \geq 1\}$. More detailed examples can be found in [2] and [10].

Now we have all the tools to state the first main result of this paper:

Theorem 2.5 (Annihilation of the error in a long time horizon) *Let* $\mathbf{x} \in (\mathbb{R}^d)^N_*$, σ *be a Lipschitz activation function,* $\mathcal{E}$ *be an error function such that Hypotheses 1 is satisfied and* J_T *given in* (2.1). *Then, for* $T > 0$ *large enough depending on* σ, $\mathbf{x}$ *and* $\mathcal{E}$, *and for all* $\varepsilon > 0$ *there is* $\delta > 0$ *such that* $J_T(w, b) < \inf J_T + \delta$ *implies:*

$$\mathcal{E}(y(T; \mathbf{x}, w, b)) < \varepsilon. \tag{2.3}$$

Moreover, for $T > 0$ *large enough the following equality holds for any optimal trajectory:*

$$\mathcal{E}(y_T(T; \mathbf{x})) = 0. \tag{2.4}$$

Here, y *is given by* (1.2), *where we consider the solution of* (1.1).

Theorem 2.5 is proved by showing that if T is sufficiently large and if $\mathcal{E}(y(T; \mathbf{x}, w, b))$ is small and strictly positive, we can construct with the second item of Hypothesis 1 a control $(\tilde{w}, \tilde{b})$ such that:

$$J_T(\tilde{w}, \tilde{b}) \leq J_T(w, b) - \frac{1}{2}\mathcal{E}(y(T; \mathbf{x}, w, b)).$$

Their proof can be found in [1]. There, we show that the trajectories may be preserved when we perform a diffeomorphism in the time variable. Then, given a control with a non-constant norm we construct a more efficient one and we use this to construct a control for which the value of the empirical risk minimization functional is smaller for all controls with a non-constant norm.

The construction of such control is far from trivial and, as the counterexample $d = N = 1$, $\mathcal{E}(x) = x^2$ and $\sigma(s) = s$ shows, the hypotheses are rather sharp. As explained in the first part of the introduction, Theorem 2.5 improves the results presented in [2], where the authors prove that the error of the final state of the optimal trajectory is of size $\mathcal{O}(1/T)$.

2.2. Optimal trajectories for neural ODE with a homogeneous activation function

In this section we present the analogous results to those in Section 2.1 for the neural ODE (1.3) with activation functions which satisfy (1.5). Let us reformulate Hypothesis 1 in the context of (1.3):

Hypothesis 2 *Let* $\mathbf{x} \in (\mathbb{R}^d)^N_*$, *let* $\mathcal{E} : (\mathbb{R}^d)^N \mapsto \mathbb{R}^+ := [0, \infty)$ *be a continuous function, and let* y *denote* (1.6), *where we consider the solutions of* (1.3). *Then:*

1. *For the data set* $\mathbf{x}$ *there are controls:*

$$(r_*, w_*, b_*) \in L^2(0, 1; \mathcal{U}),$$

such that $\mathcal{E}(y(1; \mathbf{x}, r_*, w_*, b_*)) = 0$.

2. *There are* $C, \tilde{\varepsilon} > 0$ *both just depending on* $\mathcal{E}$ *such that for all* $\overline{\mathbf{x}} = (\overline{x}^1, \dots, \overline{x}^N) \in (\mathbb{R}^d)^N_*$ *satisfying* $\mathcal{E}(\overline{\mathbf{x}}) < \tilde{\varepsilon}$, *there are some controls* (r, w, b) *satisfying:*

$$\|(w, b)\|_{L^\infty(0,1;\mathcal{U})} < C\mathcal{E}(\overline{\mathbf{x}}),$$

such that:

$$\mathcal{E}(y(1; \overline{\mathbf{x}}, r, w, b)) = 0.$$

Example 2.6 (Hypothesis 2 in a context of ensemble controllability) Hypothesis 2 can be considered in an ensemble controllability problem. Let $\mathbf{x} \in (\mathbb{R}^d)^N_*$ for $d \geq 2$, X given in (1.4):

$$\sigma(x) = (\max\{x_1, 0\}, \dots, \max\{x_d, 0\}), \tag{2.5}$$

the activation function, $\mathbf{z} = (z^1, \dots, z^N) \in (\mathbb{R}^d)^N_*$ the targets, and $\mathcal{E}$ given by (2.2) for $E_i(x) = |x - z^i|$ the error function. Note that σ satisfies:

$$|\sigma(u)| \leq |u| \quad \forall u \in \mathbb{R}^d. \tag{2.6}$$

It can be proved that Hypothesis 2 is satisfied, being the main ideas in [10], and the complete proof in [1].

Again, we seek to get sufficient conditions so that the optimal trajectories induced by:

$$\tilde{J}_T(r, w, b) := \mathcal{E}(y(T; \mathbf{x}, r, w, b)) + \int_0^T |(w(t), b(t))|^2 dt, \tag{2.7}$$

satisfy $\mathcal{E}(y_T(T; \mathbf{x})) = 0$. Since $|r|$ is constant (see (1.4)), it makes no sense to include it in the definition of $\tilde{J}_T$. For the functional $\tilde{J}_T$ the following result holds:

Theorem 2.7 (Annihilation of the error for a sufficiently large time) *Let* σ *be a Lipschitz activation function satisfying* (1.5) *and* $\mathcal{E}$ *an error function satisfying Hypothesis 2. Then, for* $T > 0$ *large enough depending on* σ, $\mathbf{x}$ *and* $\mathcal{E}$, *and all* $\varepsilon > 0$ *there is* $\delta > 0$ *such that if* $J_T(r, w, b) < \inf J_T + \delta$:

$$\mathcal{E}(y(T; \mathbf{x}, r, w, b)) < \varepsilon. \tag{2.8}$$

Moreover, if T *is large enough and if* $\tilde{J}_T$ *has an optimal trajectory:*

$$\mathcal{E}(y_T(T; \mathbf{x})) = 0. \tag{2.9}$$

Here y *is given by* (1.2), *where we consider the solution of* (1.1).

The proof of Theorem 2.7 is analogous to that of Theorem 2.5. As with Theorem 2.5, Theorem 2.7 improves the results presented in [2], where the authors prove that the error of the optimal trajectory at a final time T is of magnitude $\mathcal{O}(1/T)$ also for the solutions of (1.3) with an activation functions satisfying (1.5).

Remark 2.8 (Existence of minimizers of $\tilde{J}_T$) We have stated "if $\tilde{J}_T$ has an optimal trajectory" in Theorem 2.7 because, as far as we know, it is an open question to see if $\tilde{J}_T$ admits a minimizer. The main obstacle to adapt the proof of Proposition 2.2 is that nonlinear functions and weak limits may not commute. However, we can improve Theorem 2.7 and obtain that for T large enough and all $\varepsilon > 0$ there are controls (r, w, b) such that $J_T(r, w, b) < \inf J_T + \varepsilon$ and $\mathcal{E}(y(T; \mathbf{x}, r, w, b)) = 0$.

Remark 2.9 (Functionals allowing expensive controls) As in [2], we can consider the functional:

$$J_{T,\delta}(w,b) := \mathcal{E}(y(T;\mathbf{x},w,b)) + \delta \int_0^T |(w(t),b(t))|^2 dt,$$

instead of J_T for (1.1), and:

$$J_{T,\delta}(r,w,b) := \mathcal{E}(y(T;\mathbf{x},r,w,b)) + \delta \int_0^T |(w(t),b(t))|^2 dt,$$

instead of J_T for (1.3)-(1.5). By linearity, it holds that:

$$J_{T,\delta}(w,b) = J_{T\delta^{-1},1}(\delta w(t\delta), \delta b(t\delta)),$$

and:

$$\tilde{J}_{T,\delta}(r,w,b) = \tilde{J}_{T\delta^{-1},1}(r(t\delta), \delta w(t\delta), \delta b(t\delta)),$$

respectively. A straight consequence is that (w,b) is a minimizer of $J_{T,\delta}$ if and only if $(\delta w(t\delta), \delta b(t\delta))$ is a minimizer of $J_{T\delta^{-1},1}$. Similarly, (r,w,b) is a minimizer of $\tilde{J}_{T,\delta}$ if and only if $(r(t\delta), \delta w(t\delta), \delta b(t\delta))$ is a minimizer of $J_{T\delta^{-1},1}$. Thus, analogous results to Theorems 2.5 and 2.7 and all the auxiliary results hold true for $J_{T,\delta}$ and $\tilde{J}_{T,\delta}$ when T is fixed and $\delta > 0$ is small enough depending on σ, $\mathcal{E}$, $\mathbf{x}$ and T.

3. Open problems

- **Optimal control for non-homogenous activation functions.** It remains an open problem to determine if similar results to Theorem 2.7 hold for non-homogeneous activation functions satisfying $\sigma(0) = 0$ such as the hyperbolic tangent:

  $$\sigma(x) = (\tanh(x_1), \dots, \tanh(x_d)),$$

  see [4]. We may wonder whether similar results hold with more general activation functions if we replace X (see (1.4)) by the unitary matrices or by $\mathbb{R}^{d\times d}$ (of course, the cost of r must also be included in the risk minimization functional). This would include, for instance, sigmoid:

  $$\sigma(x) = ((1+e^{-x_1})^{-1}, \dots, (1+e^{-x_d})^{-1}),$$

  see [8]; softplus:

  $$\sigma(x) = (\log(1+e^{x_1}), \dots, \log(1+e^{x_d})),$$

  see [5], and others like logistic and cross-entropy functions. The main difficulty is that changing the speed of the control is not enough, so another tool is needed to prove the main result, probably a local inverse theorem result.

- **Optimal control with the H^1 norm.** It is a relevant problem to determine if similar results to Theorems 2.5 and 2.7 hold for any other Lebesgue or Sobolev penalty. In particular, an interesting scenario is to replace both in J_T and $\tilde{J}_T$ the terms $\|(w,b)\|^2_{L^2(0,T;\mathcal{U})}$ by $\|(w,b)\|^2_{H^1(0,T;\mathcal{U})}$ and adding the restriction that the component of r can only change signs if $(w,b) = 0$ or to measure the H^1 norm of r if the space X is connected. The interest of this is double: thinking in potential applications it makes sense to also try to bound the variations in the time variable, which can be obtained by minimizing the time derivative. Moreover, if we consider the H^1-norm we can prove as in Proposition 2.2 that $\tilde{J}_T$ admits a minimizer. The main difficulty is that we cannot define the control on $[T, T+\tau]$ independently to the controls on $[0,T]$ due to the necessity of bounding the time derivative.

- **Optimal control with the BV norm.** It is also a relevant problem to determine if results similar to Theorems 2.5 and 2.7 hold when we consider a BV penalty. The existence of minimizers, as shown in [2, Section 4], follows from the fact that any minimizing sequence in BV converges strongly in L^1. However, the main difficulty when studying these penalties, as before, is to keep track of the jumps, as we cannot define the control on $[T, T+\tau]$ independently to the controls on $[0,T]$.

Acknowledgements

This article has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement NO: 694126-DyCon). It is also supported by the Grant PID2021-126813NB-I00 funded by MICIU/AEI/10.13039/501100011033 and by "ERDF A way of making Europe", and by the grant IT1615-22 funded the Basque Government.

References

[1] J.A. Bárcena-Petisco. *Simultaneous controllability of two vibrating strings with variable coefficients,* Accepted in Advances in Differential Equations.

[2] C. Esteve, B. Geshkovski, D. Pighin, and E. Zuazua. *Large-time asymptotics in deep learning,* arXiv:2008.02491v2, 2021.

[3] C. Esteve-Yagüe and B. Geshkovski. *Sparsity in long-time control of neural odes,* Syst. Control Lett., 172:105452, (2023).

[4] E. Fathi and B. M. Shoja. *Deep neural networks for natural language processing,* In Handbook of statistics, Elsevier, volume 38, (2018) 229–316.

[5] X. Glorot, A. Bordes, and Y. Bengio. *Deep sparse rectifier neural networks,* In Proceedings of the fourteenth international conference on artificial intelligence and statistics, (2011) pages 315–323. JMLR Workshop and Conference Proceedings.

[6] E. Haber and L. Ruthotto. *Stable architectures for deep neural networks,* Inverse Probl., 34(1):014004, (2017).

[7] K. He, X. Zhang, S. Ren, and J. Sun. *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,* In *Proceedings of the IEEE international conference on computer vision*, pages (2015) 1026–1034.

[8] J. Mira and F. Sandoval. *From Natural to Artificial Neural Computation: International Workshop on Artificial Neural Networks, Malaga-Torremolinos, Spain, June 7-9, 1995: Proceedings*, volume 930. Springer Science & Business Media, (1995).

[9] V. Nair and G. E. Hinton. *Rectified linear units improve restricted boltzmann machines,* In 27th International Conference on International Conference on Machine Learning, ICML 10, (2010) 807–814.

[10] D. Ruiz-Balet and E. Zuazua. *Neural ode control for classification, approximation, and transport, SIAM Review*, 65(3): (2023) 735–773.

[11] E. Sontag and H. Sussmann. *Complete controllability of continuous-time recurrent neural networks,* Syst. Control Lett., 30(4): (1997) 177–183.

[12] E. D. Sontag. *Neural nets as systems models and controllers,* In Proc. Seventh Yale Workshop on Adaptive and Learning Systems, (1992) 73–79.

[13] E. Weinan. *A proposal on machine learning via dynamical systems,* Commun. Math. Stat., 5(1): (2017) 1–11.

[14] R. Zbikowski. *Lie algebra of recurrent neural networks and identifiability,* In 1993 American Control Conference, (1993) 2900–2901.

French-German-Spanish Conference on Optimization
Gijón, June 18-21, 2024
(pp. 37-45)

On the equivalence of some relaxations of optimal control problems on unbounded time domains

Ilya Dikariev[1], **Sabine Pickenhain**[2]

1. dikarill@b-tu.de Brandenburg Technical University Cottbus-Senftenberg, Germany
2. sabine.pickenhain@b-tu.de Brandenburg Technical University Cottbus-Senftenberg, Germany

1. Introduction

Relaxation methods are a general concept for solving problems that lack convexity. There are several such methods, and we consider three of them: Γ-regularization by [3], Young measures by [4], and convex combinations by [2]. For bounded time domains, the comparisons are mostly done by Roubíček in [8], considering different generalizations of Young measures.

We consider the relaxations for unbounded time domains and/or unbounded control sets. We establish sufficient conditions under which all these three types of relaxations are equivalent to each other. Furthermore, we give an example showing that in some cases the relaxations differ.

The equivalence to the relaxation of the problem via convex combinations is convenient for computations. This type of formulation does not introduce any new mathematical objects such as Radon measures or bipolars, but rather involves no more that functions, derivatives, and so on.

In the scenario where two problems (PR_1), (PR_2) are equivalent, one can establish the existence of an optimal solution for the first by proving the existence for the other, and vice versa. In the subject "Existence Theorem for Relaxed Control Problems on Infinite Time Horizon Utilizing Weight Functions" on the conference (FGS2024, Gijón), we present existence results for relaxed optimal control problems utilizing Young measures technique. In this manner, one can automatically derive existence results for other equivalent relaxations.

In the following, we present only the proofs that are not contained in the cited works, or that need modification.

Definition 1.1 Let (P_1), (P_2) be two abstract optimization problems with admissible sets A_1, A_2 and real valued objectives J_1, J_2:

$$\begin{array}{ll} & J_1(x) \longrightarrow Min \\ s.t. & x \in A_1 \end{array} \quad (\mathsf{P}_1), \qquad \begin{array}{ll} & J_2(y) \longrightarrow Min \\ s.t. & y \in A_2 \end{array} \quad (\mathsf{P}_2).$$

We call the problems (P_1), (P_2) **equivalent** if there are two mappings $\iota_1 : A_1 \to A_2, \iota_2 : A_2 \to A_1$ with the property $J_2(\iota_1(x)) \leq J_1(x)$ (resp. $J_1(\iota_2(y)) \leq J_2(y)$) for all $x \in A_1$ (resp. $y \in A_2$).

It follows from this definition that the mappings $\iota_{1,2}$ map minimizing sequences (optimal solution) of J_1 to minimizing sequences (optimal solution) of J_2 and vice versa.

Lemma 1.2 *Let the problems (P_1), (P_2) be equivalent with corresponding mappings ι_1, ι_2. Furthermore, let $\{x_i\}_{i\in\mathbb{N}}$ be a minimizing sequence of $J_1(x)$. Then $\iota_1(x_i)$ represents a minimizing sequence of $J_2(x)$. Moreover, if x^* is an optimal solution of (P_1), then $\iota_1(x^*)$ forms an optimal solution of (P_2).*

Proof *We denote as y_i the images $\iota_1(x_i)$ and assume that there exists $\bar{y} \in A_2$ with $J_2(\bar{y}) < \inf\limits_{i\in\mathbb{N}} J_2(y_i)$. We then obtain a contradiction to $\{x_i\}$ being a minimizing sequence because the image $\iota_2(\bar{y})$ is admissible for (P_1), i.e. lies in A_1, and*

$$\forall i \in \mathbb{N}: \quad J_1(\iota_2(\bar{y})) \leq J_2(\bar{y}) < J_2(y_i) \leq J_1(x_i).$$

The second statement is rather trivial. One consider the existence of an admissible solution $\bar{y} \in A_2$ with $J_2(\bar{y}) < J_2(\iota_1(x^*))$, and we obtain a contradiction to $J_1(x^*) = \inf\limits_{x\in A_1} J_1(x)$:

$$J_1(\iota_2(\bar{y})) \leq J_2(\bar{y}) < J_2(\iota_1(x^*)) \leq J_1(x^*).$$

□

 ISBN 978-84-10135-30-7

We relax an optimal control problem of following type:

$$\begin{aligned} J(x,u) &= \int_\Omega r(t,x(t),u(t))\mathrm{d}t \;\longrightarrow\; Min, \\ \dot{x}(t) &= f(t,x(t),u(t)) \text{ a.e. on } \Omega,\ x(t_0)=x_0, \\ x &\in W_p^{1,n}(\Omega,v), \\ u(t) &\in U \subseteq \mathbb{R}^m \text{ a.e. on } \Omega, \end{aligned} \tag{P}$$

where $f(t,\xi,\mathrm{v})$ is a Carathéodory function $\Omega \times \mathbb{R}^{n+m} \to \mathbb{R}^n$, $r(t,\xi,\mathrm{v})$ is a real valued normal integrand $\Omega \times \mathbb{R}^{n+m} \to \mathbb{R}$, and U is some closed set of $\mathbb{R}^m$. We call a variable v the **control variable** and ξ the **state variable**. We use weighted Sobolev spaces as a state space, and the weight v and the exponent p are supposed to be chosen in a way that $W_p^{1,n}(\Omega,v)$ forms a Banach space and such that for every element $x \in W_p^{1,n}(\Omega,v)$ there exists an absolutely continuous representative[1]. In following, we do not distinguish between elements from $x \in W_p^{1,n}(\Omega,v)$ and their absolute continuous representatives.

2. Preliminaries

Let us start with some definitions from [7] and [3]. Let X be a set from a Euclidean space of finite dimension, and let $\Omega \subseteq \mathbb{R}$ be an open set. Furthermore we utilize following conventions

$$\sup \emptyset = -\infty,\ \inf \emptyset = +\infty.$$

Moreover, we denote the convex hull and the closed convex hull of some set A by $\operatorname{co} A$ and $\overline{\operatorname{co}}\, A$ resp.

Definition 2.1 The function $g : \Omega \times X \to \overline{\mathbb{R}}$ is a **normal integrand** if

1. $g(t,\cdot) : X \to \overline{\mathbb{R}}$ is a l.s.c. function for a.a. $t \in \Omega$,
2. there exists a measurable function $\tilde{g} : \Omega \times X \to \overline{\mathbb{R}}$ such that $\tilde{g}(t,\cdot) = g(t,\cdot)$ for a.a. $t \in \Omega$.

Definition 2.2 The function $g : \Omega \times X \to \mathbb{R}$ is a **Carathéodory function** if

1. $g(t,\cdot) : X \to \mathbb{R}$ is a continuous function for a.a. $t \in \Omega$,
2. there exists a measurable function $\tilde{g} : \Omega \times X \to \mathbb{R}$ such that $\tilde{g}(t,\cdot) = g(t,\cdot)$ for a.a. $t \in \Omega$.

Lemma 2.3 *Let $g : \Omega \times (\mathbb{R}^n \times \mathbb{R}^l) \to \overline{\mathbb{R}}, (t,\xi,\mathrm{v}) \mapsto g(t,\xi,\mathrm{v})$ be some normal integrand and x some measurable mapping $\Omega \to \mathbb{R}^n$. Then the function $g \circ x$ defined as $g \circ x : (t,\mathrm{v}) \mapsto g(t,x(t),\mathrm{v})$ is a normal integrand on $\Omega \times \mathbb{R}^l$. In this sense we can identify*

Proof *Follows immediately from [7, Cor.2B].* □

Definition 2.4 Let $\Gamma : \Omega \to \mathcal{P}(X)$ be some set valued mapping. We call Γ **measurable** if for every closed set $A \subset X$ the set

$$\Gamma^{-1}(A) := \left\{ t \in \Omega \,\middle|\, \Gamma(t) \cap A \neq \emptyset \right\}$$

is measurable. We call Γ **closed-valued** if for every $t \in \Omega$ the set $\Gamma(t)$ is closed. Further we define $\operatorname{dom} \Gamma := \left\{ t \in \Omega \,\middle|\, \Gamma(t) \neq \emptyset \right\}$.

Lemma 2.5 *For a measurable closed valued multifunction $\Gamma : \Omega \to \mathcal{P}(\mathbb{R}^n)$ there exists at least one **measurable selection**, i.e. a function $u : \operatorname{dom} \Gamma \to \mathbb{R}^n$ with $u(t) \in \Gamma(t)$ for all $t \in \operatorname{dom} \Gamma$.*

Now we introduce the Γ-regularization (see [3, p.14]).

Definition 2.6 Let Y be a real convex space, and $g : Y \to \overline{\mathbb{R}}$. We call a pointwise supremum of continuous affine functions $Y \to \mathbb{R}$, that are everywhere less than g, a **Γ-regularization** g^{**} of g[2].

The Γ-regularization is always l.s.c. and convex, [3, Prop.3.1.].

Now we cite a sufficient condition for the invariance of a normal integrand property under Γ-regularization ([3, p.246, Prop.2.1]).

[1] See [5, 6].

[2] As g^{**} we denote a bipolar of g, which for local convex spaces coincides with Γ-regularization, [3].

Lemma 2.7 *Let $g(t,\xi,\mathrm{v})$ be a normal integrand on $\Omega \times \mathbb{R}^{n+l}$ and satisfies $\Phi(\|\mathrm{v}\|) \leq g(t,\xi,\mathrm{v})$, where the function $\Phi : [0,\infty) \to \overline{\mathbb{R}}$ is convex, increasing, l.s.c. and fulfills $\lim\limits_{z\to\infty} \frac{\Phi(z)}{z} = +\infty$. Then the Γ-regularization $g^{**}(t,\xi,\mathrm{v})$ is a normal integrand on $\Omega \times \mathbb{R}^{n+l}$ and satisfies $\Phi(\|\mathrm{v}\|) \leq g^{**}(t,\xi,\mathrm{v})$.*

Lemma 2.8 *The integrand $g(t,\xi,\mathrm{v})$ is normal iff $\chi_K(t)g(t,\xi,\mathrm{v})$ is normal for every $K \in \mathrm{comp}(\Omega)$.*

Proof *One direction of this statement is obvious. For the other one we remark that the supremum $g(t,\xi,\mathrm{v}) = \sup_{i\in J} g_i(t,\xi,\mathrm{v})$ over some countable family J of normal integrands is normal, [7, Prop.2L]. Since $\mathbb{R}$ is the union of countably many compact subsets the statement of the lemma follows immediately.* □

3. Equivalence of Γ-regularization and Convex combinations

Let us define a relaxation of a problem (P) in the sense of Γ-regularization (PRG) and in the sense of convex combinations (PRC).

$$\begin{aligned} &J_{(\mathrm{PRG})}(x) = \int_\Omega g^{**}(t,x(t),\dot{x}(t))\mathrm{d}t \to Min,\\ &g(t,\xi,\eta) = \inf\left\{r(t,\xi,\mathrm{v}) \,\middle|\, \mathrm{v} \in U \subseteq \mathbb{R}^m, f(t,\xi,\mathrm{v}) = \eta\right\},\\ &\qquad x \in W_p^{1,n}(\Omega,\nu),\ x(t_0) = x_0. \end{aligned} \tag{PRG}$$

The Γ-regularization g^{**} is obtained from g resp. to variable η. It follows from the definition of g in (PRG) that the function $g(t,x(t),\dot{x}(t))$ takes the value $+\infty$ for every t with $\forall \mathrm{v} \in U : \dot{x}(t) \neq f(t,x(t),\mathrm{v})$. Thus, we know that for any admissible solution x, the set

$$\left\{t \in \Omega \,\middle|\, \forall \mathrm{v} \in U : \dot{x}(t) \neq f(t,x(t),\mathrm{v})\right\}$$

forms a negligible set (set of measure zero).

$$\begin{aligned} &J_{(\mathrm{PRC})}(x,\lambda,u) = \int_\Omega \sum_{i=1}^{n+1} \lambda_i(t) r(t,x(t),u_i(t))\mathrm{d}t \to Min,\\ &\dot{x}(t) = \sum_{i=1}^{n+1} \lambda_i(t) f(t,x(t),u_i(t)) \text{ a.e. on } \Omega,\ x(t_0) = x_0,\\ &x \in W_p^{1,n}(\Omega,\nu),\\ &\lambda(t) \in E^n := \mathrm{co}\{e_1,\dots,e_{n+1}\} \text{ a.e. on } \Omega,\\ &u_i(t) \in U \subseteq \mathbb{R}^m \text{ a.e. on } \Omega,\\ &u_i, \lambda_i - \text{measurable for } i = 1 \dots n+1. \end{aligned} \tag{PRC}$$

Notice that the set E^n is n-dimensional, being the convex hull of $n+1$ points of dimension n.

From now on we define the function $\Psi(t,z) : \Omega \times [0,+\infty) \to \overline{\mathbb{R}}$ as a non-decreasing, convex l.s.c. in z function with the property

$$\lim_{z\to\infty} \frac{\Psi(t,z)}{z} = +\infty \text{ uniformly on every } K \in \mathrm{comp}(\Omega). \tag{C}$$

The integrand r satisfies a growth condition (G) if holds

$$\Psi(t,\|\mathrm{v}\|) \leq r(t,\xi,\mathrm{v}) \tag{G}$$

with Ψ satisfying (C).

Lemma 3.1 *We consider the problem* (PRG). *Let the integrand $r(t,\xi,\mathrm{v})$ satisfy growth condition* (G). *Let the function f be a Carathéodory-function, and U be a closed set. Then the functions $g(t,\xi,\eta)$ and its Γ-regularization $g^{**}(t,\xi,\eta)$ are normal integrands on $\Omega \times \mathbb{R}^{2n}$.*

Proof *Let K be some compact subset of Ω. We use a variant of Scorzà-Dragoni theorem for normal integrands, [3, Thm.1.1]. We show that*

$$\forall \varepsilon > 0\ \exists K_\varepsilon \subset K : |K \setminus K_\varepsilon| \leq \varepsilon \text{ and } g\big|_{K_\varepsilon \times \mathbb{R}^{2n}} \text{ l.s.c.} \tag{3.3}$$

Since r is a normal integrand we can establish condition (3.3) *for $r(t,\xi,\mathrm{v})$ restricted to $K_\varepsilon \times \mathbb{R}^{n+m}$, instead of g.*

We consider some sequence $\{(t_i,\xi_i,\eta_i)\} \subset K_\varepsilon \times \mathbb{R}^{2n}$ converging to $(\bar{t},\bar{\xi},\bar{\eta})$ and show $g(\bar{t},\bar{\xi},\bar{\eta}) \le \varliminf_{i\to\infty} g(t_i,\xi_i,\eta_i)$. We only need to show the inequality for the case that the limes inferior is a real number from $[0,\infty)$. We take a subsequence, that represents the limes inferior. For simplicity let the sequence be again $\{(t_i,\xi_i,\eta_i)\}$ and we have

$$\lim_{i\to\infty} g(t_i,\xi_i,\eta_i) = \alpha < +\infty. \tag{3.4}$$

For sufficiently large indexes i we have $g(t_i,\xi_i,\eta_i) < +\infty$, which means

$$\left\{ r(t_i,\xi_i,\mathrm{v}) \,\middle|\, \mathrm{v} \in U, f(t_i,\xi_i,\mathrm{v}) = \eta_i \right\} \neq \emptyset.$$

For every (t_i,ξ_i) the level sets of $r(t_i,\xi_i,\cdot) : U \to \mathbb{R}$ are compact since we have $\Psi(t_i,\|\mathrm{v}\|) \le r(t_i,\xi_i,\mathrm{v})$ and the function $\Psi(t,\|\mathrm{v}\|)$ fulfills (C). *Since the function r is l.s.c. in v and the preimage $f^{-1}(t,\xi,\cdot)(\eta_i)$ is closed we obtain for every (t_i,ξ_i,η_i) a $\mathrm{v}_i \in U$ with*

$$g(t_i,\xi_i,\eta_i) = r(t_i,\xi_i,\mathrm{v}_i) \text{ and } f(t_i,\xi_i,\mathrm{v}_i) = \eta_i.$$

Again in view of (C) *we obtain that all of v_i lie in some compact subset of U, and finally we obtain a subsequence $(t_{i_j},\xi_{i_j},\mathrm{v}_{i_j})$ converging to $(\bar{t},\bar{\xi},\bar{\mathrm{v}})$ and in view of continuity of f and l.s.c. of r on $K_\varepsilon \times \mathbb{R}^{n+m}$ we have*

$$\begin{aligned} f(\bar{t},\bar{\xi},\bar{\mathrm{v}}) &= \bar{\eta}, \\ r(\bar{t},\bar{\xi},\bar{\mathrm{v}}) &\le \varliminf_{j\to\infty} r(t_{i_j},\xi_{i_j},\mathrm{v}_{i_j}). \end{aligned}$$

From latter inequality and definition of g we obtain

$$g(\bar{t},\bar{\xi},\bar{\eta}) \le r(\bar{t},\bar{\xi},\bar{\mathrm{v}}) \le \varliminf_{j\to\infty} r(t_{i_j},\xi_{i_j},\mathrm{v}_{i_j}) = \varliminf_{j\to\infty} g(t_{i_j},\xi_{i_j},\eta_{i_j}) = \alpha.$$

The last limes inferior is equal to α because of (3.4). *Thus, we obtain that $g(t,\xi,\mathrm{v})$ is a normal integrand on $\Omega\times\mathbb{R}^{n+m}$. Finally, lemma 2.8 together with lemma 2.7 deliver that $g^{**}(t,\xi,\mathrm{v})$ is a normal integrand on $\Omega\times\mathbb{R}^{n+m}$ as well.* □

Lemma 3.2 *Let the integrand r satisfy growth condition* (G). *Moreover, let $x : \Omega \to \mathbb{R}^n, y : \Omega \to \mathbb{R}^n$ be measurable. Then there exist $n+1$ measurable functions $y_i : \Omega \to \mathbb{R}^n, i = 1 \dots n+1$ and $\lambda : \Omega \to E^n$, such that we have for almost every $t \in \Omega$:*

$$\begin{aligned} g^{**}(t,x(t),y(t)) &= \sum_{i=1}^{n+1} \lambda_i(t) g(t,x(t),y_i(t)), \\ y(t) &= \sum_{i=1}^{n+1} \lambda_i(t) y_i(t). \end{aligned} \tag{3.5}$$

Proof *From lemma 3.1 follows that $g^{**}(t,\xi,\eta)$ is a normal integrand on $\Omega \times \mathbb{R}^{2n}$ and corollary [7, Cor.2B] delivers that $g^{**}(t,x(t),\eta)$ and $g(t,x(t),\eta)$ are both normal integrands on $\Omega\times\mathbb{R}^n$, and due to [3, Prop.3.1.] we obtain representation* (3.5). □

Lemma 3.3 *Let x be an admissible solution of* (PRG), *and the integrand r satisfy growth condition* (G). *Then there exist functions $u : \Omega \to U^{n+1}$ and $\lambda : \Omega \to E^n$ such that the triple (x,λ,u) is admissible for* (PRC) *and $J_{(\mathrm{PRG})}(x) = J_{(\mathrm{PRC})}(x,\lambda,u)$.*

Proof *From lemma 3.2 we obtain measurable functions $\lambda_i(t), y_i(t), i = 1 \dots n+1$, which fulfill*

$$\begin{aligned} g^{**}(t,x(t),\dot{x}(t)) &= \sum_{i=1}^{n+1} \lambda_i(t) g(t,x(t),y_i(t)), \\ \dot{x}(t) &= \sum_{i=1}^{n+1} \lambda_i(t) y_i(t). \end{aligned}$$

Now we need to define a proper selection u_i, for every function y_i, to fulfill the state equation

$$\dot{x}(t) = \sum_{i=1}^{n+1} \lambda_i(t) f(t, x(t), u_i(t)).$$

For every $y_i(t)$ we define a set valued mapping

$$\Gamma_i(t) := \left\{ v \in U \,\middle|\, r(t, x(t), v) = g(t, x(t), y_i(t)) \right\}. \tag{3.6}$$

The function $g(t, x(t), y_i(t)) : \Omega \to \overline{\mathbb{R}}$ is measurable (lemma 2.3) and by [7, Thm.2J] we obtain that $\Gamma_i(t)$ are measurable set valued mappings with closed values, and for every t with $\Gamma_i(t) \neq \emptyset$ (follows from growth condition (G) *as in proof of lemma 3.1). That is the case for every $t \in \operatorname{dom} \Gamma_i$ because of coercivity of Ψ in z (see the proof of lemma 3.1). The set $\Omega \setminus \bigcap_{i=1}^{n+1} \operatorname{dom} \Gamma_i$ is negligible, because $\bar{x}$ is an admissible solution with $J(\bar{x}) < +\infty$. The same theorem [7, Thm.2J] delivers that there exists a measurable selection $u_i(t)$ for every i such that $u_i(t) \in \Gamma_i(t)$ and $y_i(t) = f(t, x(t), u_i(t))$ for all $t \in \operatorname{dom} \Gamma_i(t)$. And finally using* (3.6) *we get:*

$$g^{**}(t, x(t), \dot{x}(t)) = \sum_{i=1}^{n+1} \lambda_i(t) r(t, x(t), u_i(t)), \tag{3.7}$$

$$\dot{x}(t) = \sum_{i=1}^{n+1} \lambda_i(t) f(t, x(t), u_i(t))$$

for almost all $t \in \Omega$. The solution (x, λ, u) with $\lambda = (\lambda_1, \dots, \lambda_{n+1}), u = (u_1, \dots, u_{n+1})$ is then an admissible solution of (PRC) *and, because of* (3.7)*, we have $J_{(\mathrm{PRC})}(x, \lambda, u) = J_{(\mathrm{PRG})}$.* □

Lemma 3.4 *Let (x, λ, u) be an admissible solution of* (PRC) *and the integrand r satisfy the growth condition* (G)*. Then x is an admissible solution of* (PRG) *and $J_{(\mathrm{PRG})}(x) \leq J_{(\mathrm{PRC})}(x, \lambda, u)$.*

Proof *From the definition of function g in* (PRG) *we obtain*

$$g(t, x(t), f(t, x(t), u_i(t))) \leq r(t, x(t), u_i(t)) \text{ a.e.}$$

We make use of [3, Lemma 3.3.] and get

$$g^{**}(t, x(t), \dot{x}(t)) \leq \sum_{i=1}^{n+1} \lambda_i(t) g\big(t, x(t), f(t, x(t), u_i(t))\big) \leq \sum_{i=1}^{n+1} \lambda_i(t) r(t, x(t), u_i(t))$$

with $\lambda(t) \in E^n$ a.e. on Ω. □

Lemmas 3.4 and 3.3 imply immediately the equivalence of problems (PRC) and (PRG) in the sense of definition 1.1.

4. Equivalence of Young measures and Convex combinations

We first extend the notion of Young measure, as stated in [4], to unbounded domains Ω and sets U, which are closed, but not necessarily bounded.

Definition 4.1 We call a family of Radon measures[3] $\mu = \{\mu_t\}_{t \in \Omega}$ on U a **generalized control** and write $\mu \in \mathcal{M}_U$ if it fulfills:

i) $\operatorname{supp} \mu_t \subseteq U$ for almost all $t \in \Omega$,

ii) μ_t is a probability measure for almost all $t \in \Omega$,

iii) for every $g \in C_c(\Omega \times U)$ the function

$$h(t) = \langle \mu_t, g(t, v) \rangle := \int_U g(t, v) \mathrm{d}\mu_t(v)$$

is measurable.

[3] For the theory of Radon measures we refer to [1].

Now we are ready to define a relaxation in the sense of Young measures (also known as Gamkrelidze controls):

$$\begin{aligned} J_{(\mathrm{PRY})}(x,\mu) &= \int_\Omega \langle \mu_t, r(t,x(t),\mathrm{v})\rangle \,\mathrm{d}t \;\longrightarrow\; Min, \\ \dot{x}(t) &= \langle \mu_t, f(t,x(t),\mathrm{v})\rangle \text{ a.e. on } \Omega,\ x(t_0)=x_0, \\ x &\in W_p^{1,n}(\Omega,v), \\ \mu &\in \mathcal{M}_U. \end{aligned} \tag{PRY}$$

For further explanations we need following two definitions of **orientor fields**

$$P(t,\xi) = \left\{ \begin{pmatrix} r(t,\xi,\mathrm{v}) \\ f(t,\xi,\mathrm{v}) \end{pmatrix} \,\middle|\, \mathrm{v}\in U \right\},$$

$$P_{\mathcal{M}}(t,\xi) = \left\{ \left\langle \hat{\mu}, \begin{pmatrix} r(t,\xi,\mathrm{v}) \\ f(t,\xi,\mathrm{v}) \end{pmatrix} \right\rangle \,\middle|\, \operatorname{supp}\hat{\mu}\subseteq U,\ \hat{\mu}\text{ – probability measure} \right\}.$$

The following lemma is a modification of [4, Assertion 2.1.].

Lemma 4.2 *Let U be some closed subset of $\mathbb{R}^m$, the function $g : U \to \mathbb{R}^n$ be continuous, and let $H^k \subset \mathbb{R}^n$ be some hyperplane of dimension k, where $1 \le k \le n$. Let the probability measure $\hat{\mu}$ on U be such that* $\operatorname{supp}\hat{\mu} \subseteq g^{-1}(H^k)$*. Further, let the point $p := \langle \hat{\mu}, g\rangle$ lie in H^k and not in* co P*, where P represents the orientor field*

$$P := \left\{ g(\mathrm{v}) \,\middle|\, \mathrm{v}\in U \right\}.$$

Then there exists a hyperplane H^{k-1} of dimension $k-1$, such that $p \in H^{k-1}$ and $\operatorname{supp}\hat{\mu} \subseteq g^{-1}(H^{k-1})$*.*

Proof *Since $g^{-1}(H^k)$ contains a support of the probability measure it is not empty. We conclude that* co $P \cap H^k$ *is convex and not empty as well. We define a $k-1$-dimensional hyperplane, denoted by $H^{k-1} \subset H^k$, that separates the point p and the set* co $P \cap H^k$*. Furthermore, p lies in H^{k-1}.*

Let $\chi(\mathrm{v})$ be the characteristic function of the preimage $g^{-1}(H^k)$:

$$\chi(\mathrm{v}) := \begin{cases} 1, & g(\mathrm{v}) \in H^{k-1} \\ 0, & g(\mathrm{v}) \notin H^{k-1} \end{cases}.$$

The preimage $g^{-1}(H^k)$ is closed, as it is the preimage of a closed set under continuous mapping. Consequently, the function $\chi : U \to \mathbb{R}$ is u.s.c.

We consider the equation

$$\langle \hat{\mu}, g(\mathrm{v}) - p\rangle = 0$$

from which we deduce

$$\langle \hat{\mu}, g(\mathrm{v}) - p\rangle = \langle \hat{\mu}, \chi(\mathrm{v})(g(\mathrm{v}) - p)\rangle + \langle \hat{\mu}, (1-\chi(\mathrm{v}))(g(\mathrm{v}) - p)\rangle = 0.$$

Let $w \in H^k$ be a vector orthogonal to H^{k-1}, and directed towards co $P(t,x) \cap H^k$*. By taking a scalar product with the above equation we obtain*

$$\left\langle \hat{\mu}, \chi(\mathrm{v}) w^T (g(\mathrm{v}) - p)\right\rangle + \left\langle \hat{\mu}, (1-\chi(\mathrm{v})) w^T (g(\mathrm{v}) - p)\right\rangle = 0. \tag{4.2}$$

For all v *with $\chi(\mathrm{v}) = 1$, the scalar product $w^T(g(\mathrm{v}) - p)$ vanishes because the points $g(\mathrm{v})$ and p lie in the hyperplane H^{k-1}, and the vector w is then orthogonal to $g(\mathrm{v}) - p$. It follows*

$$\forall \mathrm{v} \in U\ \chi(\mathrm{v}) w^T (g(\mathrm{v}) - p) = 0,$$

and together with (4.2) *we conclude*

$$\left\langle \hat{\mu}, (1-\chi(\mathrm{v})) w^T (g(\mathrm{v}) - p)\right\rangle = 0. \tag{4.3}$$

Since $g(\mathrm{v}) \in$ co P*, and w is directed toward* co $P \cap H^k$*, for any* v $\in g^{-1}(H^k \setminus H^{k-1})$ *we obtain*

$$w^T (g(\mathrm{v}) - p) > 0.$$

As for v *from* $g^{-1}(H^k \setminus H^{k-1})$ *the indicator function* χ *is equal zero we conclude*

$$\forall \mathrm{v} \in U: \; g(\mathrm{v}) \in H^k \setminus H^{k-1} \Rightarrow (1-\chi(\mathrm{v}))w^T(g(\mathrm{v})-p) > 0. \tag{4.4}$$

Now from equation (4.3) *we become*

$$\int_U (1-\chi(\mathrm{v}))w^T(g(\mathrm{v})-p)\mathrm{d}\hat{\mu} = \int_{g^{-1}(H^k)} (1-\chi(\mathrm{v}))w^T(g(\mathrm{v})-p)\mathrm{d}\hat{\mu}.$$

Since $1-\chi(\mathrm{v})$ *is l.s.c., and* $w^T(g(\mathrm{v})-p)$ *is non-negative and continuous on* $g^{-1}(H^k)$*, we deduce that the function* $\mathrm{v} \mapsto (1-\chi(\mathrm{v}))w^T(g(\mathrm{v})-p)$ *is l.s.c. on* $g^{-1}(H^k)$*. Now we use a proposition [1, Ch.IV, §2(1), Prop.3] and conclude that the integrand* $(1-\chi(\mathrm{v}))w^T(g(\mathrm{v})-p)$ *vanishes on* $\operatorname{supp}\hat{\mu}$*. Now, from inequality* (4.4) *it follows that* $\operatorname{supp}\hat{\mu} \cap g^{-1}(H^k \setminus H^{k-1}) = \emptyset$*. As we assumed that* $\operatorname{supp}\hat{\mu}$ *lies in* $g^{-1}(H^k)$*, we get*

$$\operatorname{supp}\hat{\mu} \subseteq g^{-1}(H^{k-1}).$$

□

Lemma 4.3 *Let* r, f *be Carathéodory functions on* $\Omega \times \mathbb{R}^{n+m}$*, then* $\operatorname{co} P(t,\xi) = P_{\mathcal{M}}(t,\xi)$ *for almost all* $t \in \Omega$*.*

Proof *Let* (t,ξ) *be arbitrary pair from* $\Omega \times \mathbb{R}^n$ *such that* $\big(r(t,\xi,\cdot), f(t,\xi,\cdot)\big)^T : U \to \mathbb{R}^{n+1}$ *is continuous. The inclusion* $P(t,\xi) \subseteq P_{\mathcal{M}}(t,\xi)$ *is obvious. Let's show* $P_{\mathcal{M}}(t,\xi) \subseteq P(t,\xi)$*.*

Let g *be a continuous function* $g: \mathrm{v} \mapsto \big(r(t,\xi,\mathrm{v}), f(t,\xi,\mathrm{v})\big)^T$*. We assume that there exists some probability measure* $\hat{\mu}$ *with* $\langle \hat{\mu}, g\rangle = p \notin \operatorname{co} P(t,\xi)$*. Using lemma 4.2 with the settings* $k := n+1, H^k := \mathbb{R}^{n+1}$ *we obtain* $\operatorname{supp}\hat{\mu} \subseteq g^{-1}(H^n)$*, where* H^n *is a hyperplane of dimension* n*, contains the point* p*, and lies in* H^{n+1}*.*

We now set $k := n$ *and utilize the lemma 4.2 once again. After altogether* $n+1$ *repetitions we obtain that* p *lies in the hyperplane* H^0 *of dimension zero, and* $\operatorname{supp}\hat{\mu} \subseteq g^{-1}(H^0)$*. Since* $p \in H^0$ *and* $\dim H^0 = 0$ *we obtain* $H^0 = \{p\}$ *and* $g^{-1}(H^0) = g^{-1}(p)$*.*

The measure $\hat{\mu}$ *is a probability measure which implies* $\operatorname{supp}\hat{\mu} \neq \emptyset$*. Together with* $\operatorname{supp}\hat{\mu} \subseteq g^{-1}(p)$ *we obtain* $g^{-1}(p) \neq \emptyset$*, that means that there exists* $\mathrm{v} \in U$ *with* $P(t,\xi) \ni g(\mathrm{v}) = p$*, and we get a contradiction.*

□

Lemma 4.4 *Let* r, f *be Carathéodory functions on* $\Omega \times \mathbb{R}^{n+m}$ *and* (x,μ) *be an admissible solution of* (PRY)*. Then there exists an admissible solution* (x,λ,u) *of* (PRC) *with* $J_{(\mathrm{PRC})}(x,\lambda,u) \leq J_{(\mathrm{PRY})}(x,\mu)$*.*

Proof *Let us define a vector-valued function* $g: \Omega \times \mathbb{R}^m \to \mathbb{R}^{1+n}, g: (t,\mathrm{v}) \mapsto (r(t,x(t),\mathrm{v}), f(t,x(t),\mathrm{v}))^T$*. We now use lemma 4.3, and for almost all* $t \in \Omega$ *we obtain*

$$\langle \mu, g(t,\mathrm{v})\rangle = \sum_{i=1}^{n+2} \hat{\lambda}_i g(t,u_i), \quad \hat{\lambda} \in E^{n+1}, \; u_{1,\dots,n+2} \in U. \tag{4.5}$$

Now we prove that we can diminish the dimension of E^{n+1}*. We formulate following optimization problem:*

$$\begin{aligned} & c^T\tilde{\lambda} \to Min, \\ s.t. \quad & A\tilde{\lambda} = d, \\ & \tilde{\lambda} \in E^{n+1}, \end{aligned}$$

where

$$c := \begin{pmatrix} r(t,x(t),u_1) \\ \dots \\ r(t,x(t),u_{n+2}) \end{pmatrix}, \; A := \big(f(t,x(t),u_1), \dots, f(t,x(t),u_{n+2})\big), \; d := A\hat{\lambda}, \tag{4.6}$$

with v_i *and* $\hat{\lambda}$ *from* (4.5)*. Since* c *and* $\tilde{\lambda}$ *are non-negative, there exists an optimal solution* $\tilde{\lambda}^*$ *of* (4.6)*. The constraints of* (4.6) *define a convex polyhedron, therefore* $\tilde{\lambda}^*$ *lies on its boundary. It means, that there exists at least one index* $1 \leq k \leq n+2$ *with* $\tilde{\lambda}^*_k = 0$*, and it follows* $(\tilde{\lambda}^*_{i=1\dots,n+2,i\neq k}) \in E^n$*. We then obtain*

$$\text{for a.a. } t \in \Omega \; \exists \lambda \in E^n \quad \sum_{i=1}^{n+1} \lambda_i g(t,u_i) \leq \sum_{i=1}^{n+2} \hat{\lambda}_i g(t,u_i) = \langle \mu, g(t,\mathrm{v})\rangle. \tag{4.7}$$

Now we define the set-valued mapping

$$\Gamma(t) = \{(\lambda, u) \in E^n \times U^{n+1} \mid F(t,\lambda,u) = \langle \mu, f(t,x(t),\mathrm{v})\rangle, \\ F_1(t,\lambda,u) \le \langle \mu, r(t,x(t),\mathrm{v})\rangle\}$$

with $F(t,\lambda,u) = \sum_{i=1}^{n+1} \lambda_i f(t,x(t),u_i)$ *and* $F_1(t,\lambda,u) = \sum_{i=1}^{n+1} \lambda_i r(t,x(t),u_i)$. *$F$ and F_1 are Carathéodory functions. The sets $\Gamma(t)$ are not empty for a.a. $t \in \Omega$ because of* (4.7). *Theorem [7, Thm.2J] delivers that Γ is measurable, and by lemma 2.5 we get functions*

$$\lambda : \Omega \to E^n, \\ u_i : \Omega \to U,\ i = 1, \dots, n+1$$

that are measurable and $(\lambda(t), u(t)) \in \Gamma(t)$ for a.a. $t \in \Omega$. Finally, we obtain $J_{(\mathrm{PRC})}(x,\lambda,u) \le J_{(\mathrm{PRY})}(x,\mu)$. □

Lemma 4.5 *Let r be a normal integrand and f be a Carathéodory function on $\Omega \times \mathbb{R}^{n+m}$, and let (x,λ,u) be an admissible solution of* (PRC). *Then, there exists an admissible solution (x,μ) of* (PRY) *such that $J_{(\mathrm{PRY})}(x,\mu) = J_{(\mathrm{PRC})}(x,\lambda,u)$.*

Proof *The proof is straightforward: define $\mu_t := \sum_{i=1}^{n+1} \lambda_i(t)\delta_{u_i(t)}$, and it can be readily shown that $\mu := \{\mu_t\}_{t\in\Omega}$ constitutes a generalized control according to definition 4.1.* □

Now, under the more restrictive conditions of lemma 4.4 we obtain the equivalence of problems (PRY) and (PRC).

5. Example

We will now provide an example to illustrate how the relaxations differ.

$$\begin{aligned} J(x,u) &= \int_0^\infty [e^{-u^2(t)} + x^2(t)]e^{-t}\mathrm{d}t \ \longrightarrow\ Min, \\ \dot{x}(t) &= \frac{1}{1+u^2(t)},\ \text{a.e. on } (0,\infty),\ x(0) = 0, \\ x &\in W_2^1((0,\infty), e^{-t}), \\ u(t) &\in \mathbb{R} \text{ a.e. on } (0,\infty), \\ u &- \text{measurable.} \end{aligned} \tag{PEX}$$

To get the Γ-regularization we first calculate the function g according to (PRG).

$$g(t,\xi,\eta) = \inf\left\{(e^{-\mathrm{v}^2} + \xi^2)e^{-t} \mid \mathrm{v} \in \mathbb{R}, \frac{1}{1+\mathrm{v}^2} = \eta\right\} = \begin{cases} +\infty, & \eta \le 0, \\ (e^{1-\frac{1}{\eta}} + \xi^2)e^{-t}, & \eta > 0. \end{cases}$$

Now we can easily calculate the Γ-regularized function according to the definition 2.6:

$$g^{**}(t,\xi,\eta) = \begin{cases} +\infty, & \eta < 0, \\ \xi^2 e^{-t}, & \eta \ge 0. \end{cases} \tag{5.1}$$

We insert this function, g^{**}, into the formulation (PRG) and conclude that the problem

$$J_{(\mathrm{PRG})}(x) = \int_0^\infty g^{**}(t,x(t),\dot{x}(t))\mathrm{d}t \to Min, \\ x \in W_2^1((0,\infty), e^{-t}),\ x(0) = 0,$$

where the function g^{**} is taken from (5.1), possesses an optimal solution $x^* \equiv 0$ with $J_{(\mathrm{PRG})}(x^*) = 0$.

On the other hand, since the integrand $r(t,\xi,\mathrm{v}) = (e^{-\mathrm{v}^2} + \xi^2)e^{-t}$ is always greater than zero, for any probability measure $\hat{\mu}$, we obtain $\langle \hat{\mu}, (e^{-\mathrm{v}^2} + \xi^2)e^{-t}\rangle > 0$. This implies that for any generalized control μ, we have

$$J_{(\mathrm{PRY})}(x,\mu) = \int_0^\infty \langle \mu_t, e^{-\mathrm{v}^2} + x^2(t)\rangle e^{-t}\mathrm{d}t > 0.$$

At the same time, the sequence of generalized controls[4] $\mu_k := \{\delta_{kt}\}_{t\in\Omega}$ and corresponding solutions $x_k(t) := \frac{1}{k}\arctan(kt)$ of the initial value problem of (PEX) form a null sequence $J_{(\mathrm{PRY})}(x_k, \mu_k)$

$$J_{(\mathrm{PRY})}(x_k,\mu_k) = \int_0^\infty \langle \delta_{kt}, e^{-\mathrm{v}^2} + x_k^2(t)\rangle e^{-t}\mathrm{d}t = \int_0^\infty \left(e^{-k^2t^2} + \frac{1}{k^2}\arctan^2(kt)\right)e^{-t}\mathrm{d}t < \frac{\sqrt{\pi}}{2k} + \frac{\pi^2}{4k^2} \overset{k\to\infty}{\longrightarrow} 0.$$

We conclude that there is no optimal solution for either the relaxations of the type of Young measures or the convex combinations, according to lemma 4.4. Furthermore, because the condition (G) cannot be satisfied, we are unable to extract any admissible solutions for other types of relaxations discussed here from Γ-regularization.

Acknowledgments

The authors were supported by DFG Grants PI209/8-3 and LY149/2-3.

References

[1] N. Bourbaki. *Integration I: Chapters 1-6*. "Springer", 2004.

[2] D.A. Carlson. Nonconvex and relaxed infinite – horizon optimal control problems. *Journal of optimization theory and applications*, 78(3):465–491, 1993.

[3] I. Ekeland and R. Témam. *Convex Analysis and Variational Problems*. SIAM, 1999.

[4] R.V. Gamkrelidze. *Principles of Optimal Control Theory*. Plenum Press, New York and London, 1978.

[5] A. Kufner. *Weighted Sobolev Spaces*. John Wiley & Sons, 1985.

[6] A. Kufner and B. Opic. How to define reasonably weighted sobolev spaces. *Commentationes Mathematicae Universitatis Carolinae*, 025(3):537–554, 1984.

[7] R. Tyrrell Rockafellar. Integral functionals, normal integrands and measurable selections. In J.P. Gossez, E.J. Lami Dozo, J. Mawhin, and L. Waelbroeck, editors, *Lecture Notes in Mathematics: Nonlinear Operators and Calculus of Variations*, volume 543, pages 157–207. Springer, 1985.

[8] T. Roubíček. *Relaxation in Optimization Theory and Variational Calculus*. De Gruyter, Berlin, Boston, 2020.

[4] We denote by δ_x a Dirac-measure, concentrated at the point x.

French-German-Spanish Conference on Optimization
Gijón, June 18-21, 2024
(pp. 46-54)

Some questions related to geometric inverse problems

Anna Doubova
doubova@us.es Universidad de Sevilla, Dpto. Ecuaciones Diferenciales y Análisis Numérico (EDAN) and IMUS, Spain

Abstract

We will consider geometric inverse problems of determining by external measurements a portion of the domain in which certain partial differential equations are satisfied. We will consider real-world applications problems and will explore two crucial aspects: uniqueness and numerical reconstruction based on certain optimization problems. We will present results that have been obtained in collaboration with different authors.

This paper will consist of two parts. First, we will present some numerical domain reconstruction techniques related to optimization problems. We will focus on meshless technique based on the Method of Fundamental Solutions which will be introduced in the context of an elliptic equation.

The second part of the paper will be devoted to analyzing the sensitivity of the inverse problems to the boundary and initial data which we will present in the context of the variable density Burguers equation.

1. Introduction

This paper deals with geometric inverse problems for certain partial differential equations (PDEs). We aim to determine a portion of the domain where these equations hold true, based on external measurements taken on a part of the boundary. Our focus is on developing novel numerical methods of reconstruction of unknown domain and the crucial question of uniqueness. We present several results obtained through ongoing collaborations, which significantly advance our understanding of this problem.

The analysis and solution of inverse problems of many kinds has recently increased a lot because of their relevance in many applications: elastography and medical imaging, seismology, potential theory, ion transport problems or chromatography, finances, etc.; see for instance [9]. The variety of inverse problems is huge in comparison with their direct analogs and many inverse problems coming from very classical and basic direct problems wait for theoretical and numerical research. Let us mention the monographs [4, 17, 18] and [10], where many theoretical and numerical aspects of inverse problems for partial differential equations are depicted.

The paper will be structured in two parts. First, Section 2, will deal with reconstruction algorithms involving some optimization problems conceived to compute the unknown domain from boundary measurements. A meshless technique based on the method of fundamental solutions (MFS) will be used in the context of an elliptic equation.

The second part, in Section 3, will be devoted to analyze the sensitivity of inverse problems to the boundary and initial data which will be presented in the context of the variable density Burguers one dimensional equation.

The author would like to express a sincere gratitude to the collaborators of this research, particularly Jone Apraiz, Jin Cheng, Enrique Fernández-Cara, Pitágoras de Carvalho, Jairo Rocha de Faria and Masahiro Yamamoto.

2. Method of Fundamental Solutions

This section focuses on developing reconstruction algorithms for the unknown domain. We will explore optimization problems designed to compute a unknown portion of the domain from boundary measurements. In particular, we will employ a meshless technique based on the MFS within the context of an elliptic equation.

Let $\Omega \subset \mathbb{R}^N$ be a simply connected bounded open set ($N \geq 1$) whose boundary $\partial\Omega$ is of class C^2 and let γ be a nonempty open subset of $\partial\Omega$. We consider the following inverse problem:

IP-1: Given functions $\tilde{\alpha} = \tilde{\alpha}(x)$, $a = a(x)$, $h = h(x)$ and $\varphi = \varphi(x)$ in appropriate spaces, find a set D such that the solution u to the Dirichlet problem

$$\begin{cases} -\Delta u + au = h, & x \in \Omega \setminus \overline{D}, \\ u = \varphi, & x \in \partial\Omega, \\ u = 0, & x \in \partial D \end{cases} \tag{2.1}$$

 ISBN 978-84-10135-30-7

satisfies the additional condition

$$\frac{\partial u}{\partial n} = \tilde{\alpha} \quad \text{on} \quad \gamma. \tag{2.2}$$

In this context, it is usual to consider three main questions: uniqueness, stability and reconstruction. They can be described as follows.

- **Uniqueness:** Let u^1 and u^2 be solutions to (2.1) corresponding to the sets D^1 and D^2. Let us assume that the associated observations on γ coincide, that is, $\tilde{\alpha}^1 = \tilde{\alpha}^2$. Then, do we have $D^1 = D^2$?
- **Stability:** Find an estimate of the "size" of $(D^1 \setminus D^2) \cup (D^2 \setminus D^1)$ in terms of the "size" of $\tilde{\alpha}^1 - \tilde{\alpha}^2$.
- **Reconstruction:** Find an iterative algorithm to compute D from $\tilde{\alpha}$.

The uniqueness for this problem is based on the unique continuation property for the Poisson equation and can be achieved using arguments from [15]. Concerning stability, see [5].

In [7] and [8], we have considered inverse problems similar to (2.1)–(2.2) respectively for the N-dimensional wave equation and the Lamé system. We introduced some reconstruction methods based on reformulation as optimization problems and finite element techniques that require a new mesh at each iteration of the algorithm. This was implemented with the help of `FreeFem++`, see [11], used in combination with the `ff-NLopt` package.

Motivated by the fact that the identification of small obstacles is difficult and expensive with domain discretization methods and, on the other hand, trying to investigate how meshless methods work in the context of geometric inverse problems, we have used MFS.

The MFS was introduced by Kupradze and Alexidze in the 1960's (see [16]). It is an efficient meshless numerical method for the computation of solutions of linear PDEs. The key idea is to use a basis formed by fundamental solutions. Some advantages of this meshless method over classical domain discretization approach are the simplicity of implementation, the high computational speed and the exponential convergence properties, see [14]. It will be used below in combination with the method of particular solutions (MPS), see [13].

2.1. The two-dimensional case

Let us explain how MFS-MPS works for the numerical solution of (2.1)–(2.2). In order to present more clearly its application, we will consider a (geometrically simple) situation in which the unknown domain is a 2D ball.

Thus, let us assume that $N = 2$, $\Omega = B(0;R)$ (the ball centered at the origin with radius R) and $D = B(x_0;r)$ for some (unknown) x_0 and r. Let us introduce the family of admissible subdomains

$$X_b = \{(x_0;r) \in \mathbb{R}^3 \,:\, r > 0,\, \overline{B}(x_0;r) \subset \Omega\}.$$

Let us fix a non-empty open subset $\gamma \subset \partial\Omega$. Then, the inverse problem is as follows: find $(x_0,r) \in X_b$ such that the associated solution u to (2.1) satisfies (2.2). Note that, independently of the choice of (x_0,r) in X_b, for any $h \in L^2(\Omega)$ and any $\tilde{\alpha} \in H^{1/2}(\gamma)$, the solution to (2.1) belongs to $H^2(\Omega \setminus \overline{B(x_0;r)})$ and (2.2) makes sense.

The main steps of the MFS-MPS are the following:

Step 1 (MFS-MPS). Let us write the first equation from (2.1) in the form $-\Delta u = -au + h$. We look for an approximation (also denoted u) of the form $u = u_P + u_H$, where u_P is a particular solution to the non-homogeneous (complete) PDE and u_H is a solution to the Laplace equation. Specifically, we look for linear combinations of *radial basis* functions in the case of u_P and *fundamental solutions* in the case of u_H:

$$u(x) = u_P(x) + u_H(x) := \sum_{j=1}^{N_f} \beta_j F(\|x - \eta_j\|) + \sum_{k=1}^{N_b} \alpha_k G(\|x - \xi_k\|). \tag{2.3}$$

Here, we have used the following notation: N_f is the number of the field points η_j (associated to the radial functions) and N_b is number of the source points ξ_k (associated to the fundamental solutions; see Figure 1).

It will, be assumed that F is the *integrated radial basis* function, obtained by analytical integration from the equation $\Delta F = f$, where $f = f(r)$ is the so called compactly supported radial basis function (CSRBF; see [14]):

$$f(r) = \begin{cases} \left(1 - \dfrac{r}{\lambda}\right)^2 & \text{if} \quad r \le \lambda, \\ 0 & \text{if} \quad r > \lambda \end{cases} \qquad \text{and} \qquad F(r) = \begin{cases} \dfrac{r^4}{16\lambda^2} - \dfrac{2r^3}{9\lambda} + \dfrac{r^2}{4} & \text{if} \quad r \le \lambda, \\ \dfrac{13\lambda^2}{144} + \dfrac{\lambda^2}{12}\log\left(\dfrac{r}{\lambda}\right) & \text{if} \quad r > \lambda, \end{cases}$$

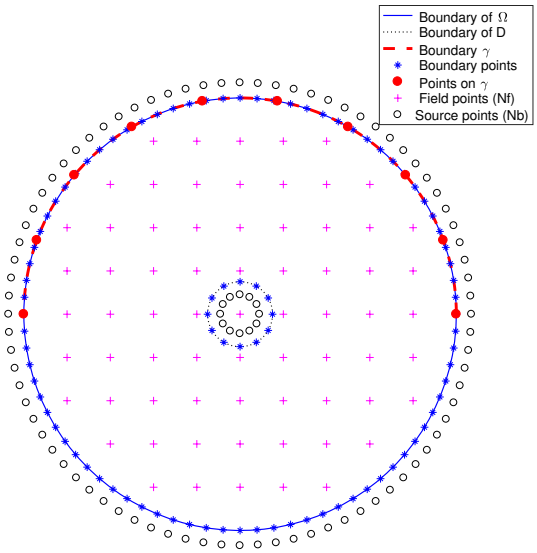


Fig. 1 Representation of a 2D domain displaying the field, source and boundary points used by the MPS-MFS.

where λ is a scaling factor. On the other hand, G is the fundamental solution of the Laplace equation. Thus,

$$G(\|x-\xi_k\|) = -\frac{1}{2\pi}\log\Big(\|x-\xi_k\|\Big),$$

where the ξ_k are the source points and $\|\cdot\|$ denotes the Euclidean norm.

Step 2 (Reduction to a nonlinear algebraic system). Considering an approximation of the solution to (2.1) of the form (2.3) and imposing (2.2), we obtain the following equations:

- *The PDE at the field points η_i:* for $i = 1, \dots, N_f$,

$$\sum_{j=1}^{N_f}\beta_j\Big[-f(\|\eta_i-\eta_j\|) + a\,F(\|\eta_i-\eta_j\|)\Big] + \sum_{k=1}^{N_b}\alpha_k\,a\,G(\|\eta_i-\xi_k\|) = h(\eta_i). \tag{2.4}$$

- *The Dirichlet boundary condition at the boundary points $\zeta_k \in \partial\Omega$:* for $m = 1, \dots, N_{b,0}$,

$$\sum_{j=1}^{N_f}\beta_j\,F(\|\zeta_m-\eta_j\|) + \sum_{k=1}^{N_b}\alpha_k\,G(\|\zeta_m-\xi_k\|) = \varphi(\zeta_m). \tag{2.5}$$

  (here, $N_{b,0}$ is the number of boundary points on $\partial\Omega$; see Figure 1).

- *The Neumann boundary condition at the boundary points $\zeta_m \in \gamma$:* for $m = 1, \dots, N_{b,1}$,

$$\sum_{j=1}^{N_f}\beta_j\,\frac{\partial}{\partial n}F(\|\zeta_m-\eta_j\|) + \sum_{k=1}^{N_b}\alpha_k\,\frac{\partial}{\partial n}G(\|\zeta_m-\xi_k\|) = \tilde{\alpha}(\zeta_m), \tag{2.6}$$

  where $N_{b,1} \le N_{b,0}$ is the number of boundary points on γ.

- *The Dirichlet boundary condition at the boundary points $d_m \in \partial D$, which are in principle unknown:* for $m = N_{b,0}+1, \dots, N_b$,

$$\sum_{j=1}^{N_f}\beta_j\,F(\|d_m-\eta_j\|) + \sum_{k=1}^{N_b}\alpha_k\,G(\|d_m-\xi_k\|) = 0. \tag{2.7}$$

  Recall that the points d_m are assumed to be located on the boundary of $D = B(x_0, r)$ for some unknown x_0 and r.

Therefore, (2.1)–(2.2) can be rewritten as the following problem for a nonlinear system of equations:

$$\begin{cases} \text{Find } (\beta,\alpha)\in\mathbb{R}^{N_f}\times\mathbb{R}^{N_b} \text{ and } (x_0,r)\in X_b \text{ such that} \\ M(x_0,r)\begin{bmatrix}\beta\\ \alpha\end{bmatrix}=Z, \end{cases} \tag{2.8}$$

where $M(x_0,r)$ and Z are found from the left and the right hand sides in (2.4)–(2.7).

Step 3 (Least squares). Let us notice that (2.8) is a nonlinear system of $N_f+N_b+N_{b,1}$ equations with N_f+N_b+3 unknowns. It possesses the following *least squares* formulation:

$$\begin{cases} \text{Find } (\beta,\alpha,x_0,r)\in X_d \text{ such that} \\ J(\beta,\alpha,x_0,r)\leq J(\beta',\alpha',x_0',r') \quad \forall\,(\beta',\alpha',x_0',r')\in X_d\,, \end{cases} \tag{2.9}$$

where

$$X_d := \mathbb{R}^{N_f}\times\mathbb{R}^{N_b}\times X_b$$

and the function $J : X_d \mapsto \mathbb{R}$ is defined by

$$J(\beta,\alpha,x_0,r) := \frac{1}{2}\left\| M(x_0,r)\begin{bmatrix}\beta\\ \alpha\end{bmatrix}-Z\right\|^2. \tag{2.10}$$

2.2. The three-dimensional case

We assume that $N=3$, $\Omega=B(0;R)$ is the ball centered at $(0,0,0)$ of radius R and $D=B(x_0;r)$ is an inner ball, centered at x_0 with radius r for some (unknown) x_0 and r.

Our aim is, again, to find a numerical approximation of the solution of the form (2.3), built with the help of integrated radial basis functions and fundamental solutions to the Laplace equation. Notice that since the spatial dimension is 3, the functions F and G are different.

If we do as before, we will have to work with too many unknowns at each iteration of the optimization algorithm. In order to avoid this difficulty, we will proceed as follows:

- We introduce a new cost functional I, only depending on x_0 and r. More precisely, given x_0, we first compute the coefficients β_j and α_k that appear in (2.3) using (2.4), (2.5) and (2.7). Then, we compute the normal derivative of the corresponding u at the points ζ_m and set

$$I(x_0,r) := \frac{1}{2}\sum_{m=1}^{N_{b,1}}\left|\frac{\partial u}{\partial n}(\zeta_m)-\tilde{\alpha}(\zeta_m)\right|^2.$$

  The field points η_i, source points ξ_k and boundary points $\zeta_k\in\partial\Omega$ and $\zeta_m\in\gamma$ are depicted in Figure 2. Again, as in the 2D case, in order to get good convergence properties, the source points must be very close to the boundary points.

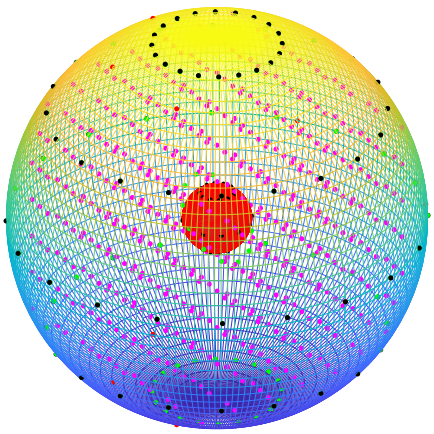


Fig. 2 Representation of a 3D domain, displaying the field points, the source points and the boundary points, as well a possible distribution of these points for MPS-MFS.

- We consider the following extremal problem:

$$\begin{cases} \text{Find } (x_0, r) \in X_s \text{ such that} \\ I(x_0, r) \leq I(x_0', r') \quad \forall\, (x_0', r') \in X_s, \end{cases} \tag{2.11}$$

where $X_s := \{\, (x_0, r) \in \mathbb{R}^4 \,:\, r > 0,\ \overline{B}(x_0; r) \subset \Omega \,\}$.

Note that the computation of $I(x_0, r)$ is more involved, since it needs the "intermediate" resolution of a system of $N_f + N_b$ equations with the same number of unknowns. However, in (2.11) the number of unknowns is just 4.

Remark 2.1 (Some open questions) There are several open questions that arise in connection with this method. It is worth highlighting the following.

1. How to explore these techniques in the case of other more complex geometries, for example polyhedral unknown D in 3D, the case of three or more balls, ...Notice that in [6] we have performed computations in the case of polygon domain and two unknown balls and 3D
2. It would be very interesting to investigate the application of MFS-MPS to the case of the evolution problems (wave equation, Lamé system, Slokes, Navier-Stokes, Boussinesq, ...).

2.3. Numerical results

For simplicity, we will present simulations only for 2D case. Let us fix $a \equiv 0.2\sqrt{x^2 + y^2}$, $h(x, y) \equiv 0.3x$ and $\varphi(x, y) \equiv 10x$. The boundary observation $\tilde{\alpha}$ has been computed from *desired* values of x_0 and r. Accordingly, our goal has been to recover these values using suitable optimization algorithms for (2.9)–(2.10).

In order to solve (2.9)–(2.10), we have performed computations using the `fmincon` function of the MATLAB© Optimization Toolbox. Let us also recall that, in order to get convergence, we must take the source points ξ_k very close to the boundary points ζ_k.

Test 1: We take $R = 10$, `x0 = (2, 4)`, `rd = 1` (the desired center and radius), `x0i =(0, 0)`, `rini = 1.5` (the center and radius of the starting ball B in the minimization algorithm). Using the MATLAB function `fmincon` with `active-set` as an optimization strategy, we find the following values (see Figure 3): `x0c = 2.000274`, `y0c = 4.000057`, `rc = 0.999658`.

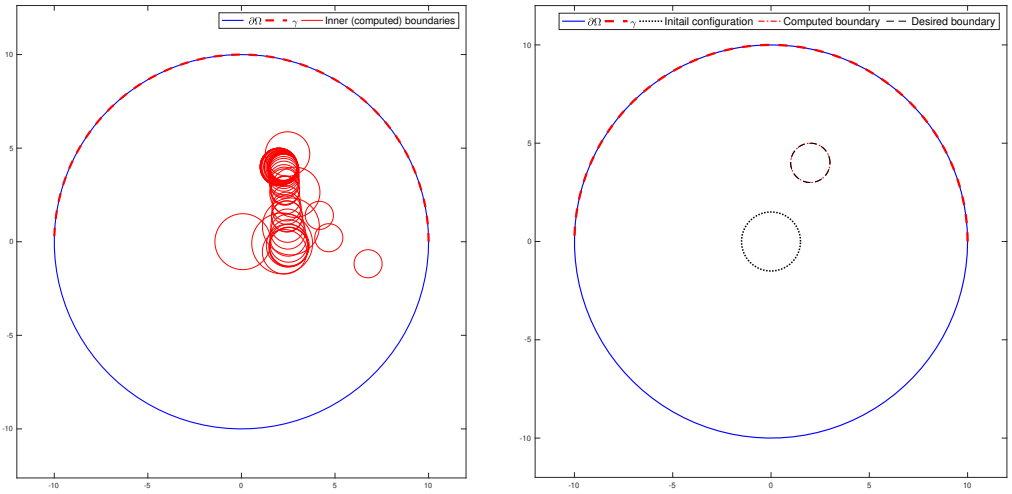


Fig. 3 Test 1 – Iterates of the optimization algorithm (left). The initial, desired and computed configurations (right). The number of iterates is 146 and the final value of the cost functional is $< 10^{-9}$. The subset γ is the part of the outer boundary marked in flashing red in dashed thick line.

Comparing with mesh depending method based on FEM, we need around 1000 iterates to get a cost $< 10^{-7}$.

3. Burgers equation and some related systems

This second part of this paper deals with a geometric inverse problem related to the identification of the size of the spatial interval where a time-dependent governing nonlinear equation must be satisfied. We will focus on the viscous non-homogeneous Burgers equation satisfied for $(x,t) \in (0,\ell) \times (0,T)$. We will assume that the equation is complemented with boundary and initial conditions corresponding to known data, respectively for $x \in \{0,\ell\}$ and $t = 0$. Then, we will try to determine the width ℓ of the spatial interval from some extra information, for instance given at $x = 0$.

The main goals will be to analyze the uniqueness (establish or discard) and to compute approximations of the solutions to the inverse problems. The details can be found in [2]. Related questions have been analyzed recently for the linear heat and wave equations in [1].

We consider a non-homogeneous (or variable density) one-dimensional fluid, modeled as follows:

$$\begin{cases} \rho(u_t + uu_x) - u_{xx} = 0, & 0 < x < \ell,\ t > 0, \\ \rho_t + u\rho_x = 0, & 0 < x < \ell,\ t > 0, \\ u(0,t) = \overline{u}(t),\ u(\ell,t) = 0, & t > 0, \\ \rho(0,t) = \overline{\rho}(t), & t \in \mathbb{R}_+ \cap \{t : \overline{u}(t) > 0\}, \\ u(x,0) = u_0(x),\ \rho(x,0) = \rho_0(x), & 0 < x < \ell. \end{cases} \tag{3.1}$$

The unknown $u = u(x,t)$ can be interpreted (for example) as the velocity of the particles of a homogeneous viscous fluid in a tube where the flow is allowed only lengthwise and $\rho = \rho(x,t)$ is the density transported with the fluid.

Of course, this can be viewed as a toy model for the variable density Navier-Stokes system. The corresponding inverse problem is the following:

IP-2: *Fix (u_0,ρ_0) and $(\overline{u},\overline{\rho})$ in (3.1) in appropriate spaces and assume that $\beta := u_x|_{x=0}$ and $\eta := \rho|_{x=0}1_{\{t:\overline{u}(t)\leq 0\}}$ are known. Then, find ℓ.*

3.1. Uniqueness

Theorem 3.1 *Assume that $0 < \ell \leq L$, $T > 0$ and (u_0,ρ_0) and $(\overline{u},\overline{\rho})$ satisfy*

$$\begin{cases} \overline{u},\overline{\rho} \in L^\infty(0,T),\ \overline{u} \not\equiv 0,\ \overline{\rho} \geq 0, \\ u_0 \equiv 0,\ \rho_0 \in L^\infty(0,L),\ \rho_0 \geq a_0 > 0. \end{cases}$$

Let (u^ℓ,ρ^ℓ) and (u^L,ρ^L) be the solutions to (3.1) for $0 < t < T$ respectively corresponding to ℓ and L. Let us assume that $|u_t^\ell| + |u_x^\ell| + |\rho_x^\ell| \leq M$ and $|u_t^L| + |u_x^L| + |\rho_x^L| \leq M$ respectively in $(0,\ell) \times (0,T)$ and $(0,L) \times (0,T)$ and $u_x^\ell(0,\cdot) = u_x^L(0,\cdot)$ and $\rho^\ell(0,\cdot) = \rho^L(0,\cdot)$. Then, $\ell = L$.

For the proof, we will use a unique continuation property satisfied by the solutions to systems of the form

$$\begin{cases} a(x,t)v_t - v_{xx} + b(x,t)v_x + c(x,t)v + d(x,t)p = 0, & (x,t) \in Q, \\ p_t + m(x,t)p_x + r(x,t)v = 0, & (x,t) \in Q, \end{cases} \tag{3.2}$$

where we assume that $Q := (0,\ell) \times (0,T)$,

$$b,c,d,m,r \in C^0(\overline{Q}), a \in C^1(\overline{Q}) \text{ and } a \geq a_0 > 0 \text{ in } Q. \tag{3.3}$$

More precisely, we have the following:

Proposition 3.2 *Assume that (3.3) is satisfied and (v,p) solves (3.2), with $v, v_x, v_{xx}, p, p_x \in C^0(\overline{Q})$. Also, assume that*

$$\begin{cases} v(0,t) = 0,\ v_x(0,t) = 0,\ p(0,t) = 0, & 0 < t < T, \\ v(x,0) = 0,\ p(x,0) = 0, & 0 < x < \ell. \end{cases} \tag{3.4}$$

Then, one has $v \equiv 0$ and $p \equiv 0$.

The proof of Proposition 3.2 can be obtained by combining two Carleman inequalities (see [2]) that can be deduced for the solutions to the first and the second equation in (3.2). The main steps are the following:

- To choose a suitable weight function (the same in both inequalities).

- To argue as in [19] and [12] and deduce appropriate estimates for v and p.
- Finally, to add and eliminate all undesirable terms on the right hand sides.

Proof of Theorem 3.1: Note that $u^\ell \in L^\infty((0,\ell)\times(0,T))$ and $u^L \in L^\infty((0,L)\times(0,T))$. If we set $v := u^\ell - u^L$ and $p := \rho^\ell - \rho^L$, one has

$$\begin{cases} \rho^\ell v_t - v_{xx} + \rho^\ell v u_x^\ell + \rho^\ell u^L v_x + (u_t^L + u^L u_x^L)p = 0, & 0 < x < \ell,\ t > 0, \\ p_t + u^L p_x + v\rho_x^\ell = 0, & 0 < x < \ell,\ t > 0, \\ v(0,t) = 0,\ \ v_x(0,t) = 0,\ \ p(0,t) = 0, & t > 0, \\ v(x,0) = 0,\ \ p(x,0) = 0, & 0 < x < \ell. \end{cases}$$

Consequently, v and p satisfies (3.2) with $a = \rho^\ell$, $b = \rho^\ell u^L$, $c = \rho^\ell u_x^\ell$, $d = u_t^L + u^L u_x^L$, $m = u^L$ and $r = \rho_x^\ell$ and (3.4).

In view of Proposition 3.2, one has $v = 0$ and $p = 0$ in $(0,\ell)\times(0,T)$. This yields $u^L(x,t) = 0$ in $(\ell, L)\times(0,T)$. Since the equations satisfied by u^L and ρ^L also possess the unique continuation property, we find that $u^L \equiv 0$, which is impossible, since $\overline{u} \not\equiv 0$. □

Remark 3.3 (Some open questions) It is worth mentioning the following open questions related to the subject:

1. It would be interesting to find nonzero initial data (u_0, ρ_0) such that uniqueness fails, in a similar way as in the case of the following viscous Burgers equation a contra-example with $u_0 \neq 0$ confirming the non-uniqueness can be found (see [2]):

$$\begin{cases} u_t - u_{xx} + uu_x = 0, & 0 < x < \ell,\ 0 < t < T, \\ u(0,t) = \eta(t),\ \ u(\ell,t) = 0, & 0 < t < T, \\ u(x,0) = u_0(x), & 0 < x < \ell. \end{cases} \tag{3.5}$$

2. On the other hand, it would also be interesting to prove a result similar to one we have for (3.5) asserting that, if the boundary data η are large enough (with respect to the other data in the system), uniqueness is satisfied. However, to our knowledge these questions for (3.1) are open.

3. In [3] we have considered similar problem for a fluids-solid interaction system. However, for other related systems, as for example viscoelastic fluids, free-boundary obstacle problem, ...these questions remain open.

3.2. Numerical results for Burgers equation

In this section, we will present a numerical experiments for the previous inverse problem. We will carry out the reconstruction of the unknown length through the resolution of some appropriate extremal problems.

The results of the numerical tests that follow will serve to illustrate the non-uniqueness result for the Burgers equation (3.5) that we have commented in the previous section.

We deal with the following

Reformulation of IP-2: *Given $T > 0$, $\eta = \eta(t)$, $u_0 = u_0(x)$ and $\beta = \beta(t)$, find $\ell \in (\ell_0, \ell_1)$ such that*

$$J_1(\ell) \leq J_1(\ell') \quad \forall\, \ell' \in (\ell_0, \ell_1), \tag{3.6}$$

where J is given by

$$J_1(\ell) := \frac{1}{2}\int_0^T |\beta(t) - u_x^\ell(0,t)|^2\, dt. \tag{3.7}$$

Here, u^ℓ is the state, i.e. the solution to (3.5) *corresponding to the length ℓ.*

Test 2: Burgers equation with $u_0 \neq 0$ and "small" η.

Here, we deal with a non-uniqueness situation. Our aim is to investigate the behavior of the algorithm in a situation of this kind.

We take $T = 6$, $\eta = 0$ in $(0,T)$ and $u_0(x) \equiv \pi\sin(\pi x/2)/(2 + \cos(\pi x/2))$. Note that we have $u_0(x) \equiv \sin(3\pi x/L_d^1)/(2 + \cos(3\pi x/L_d^1)) \equiv \sin(2\pi x/L_d^2)/(2 + \cos(2\pi x/L_d^2))$, with $L_d^1 = 6$ and $L_d^2 = 4$; consequently, this initial data can be used as in [2] to prove non-uniqueness.

We will consider the following experiments:

- First, we start from $L_i = 5.6$, and we obtain the computed value $L_c^1 = 5.998083259$ with the associated cost is $J(L_c^1) < 10^{-8}$.
- Then, we start from $L_i = 4.6$, and we obtain the computed value $L_c^2 = 4.000601673$ with the associated cost $J(L_c^2) < 10^{-9}$.

The corresponding computed boundary observations are displayed in Figures 4 and 5, respectively. Thus, we confirm that these identical observations correspond two different solutions displayed in Figures 6 and 7.

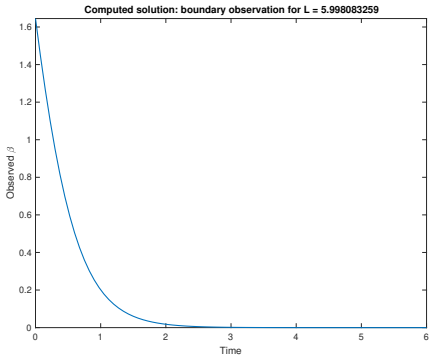


Fig. 4 Burgers equation, $\eta = 0$, fixed $u_0(x)$. The computed boundary observation $u_x(0,\cdot)$ for $L_c^1 = 5.996562049$.

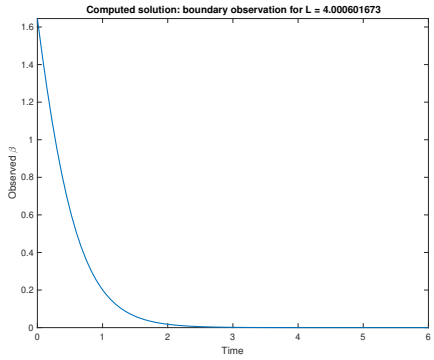


Fig. 5 Burgers equation, $\eta = 0$, fixed $u_0(x)$. The computed boundary observation $u_x(0,\cdot)$ for $L_c^2 = 4.007345905$

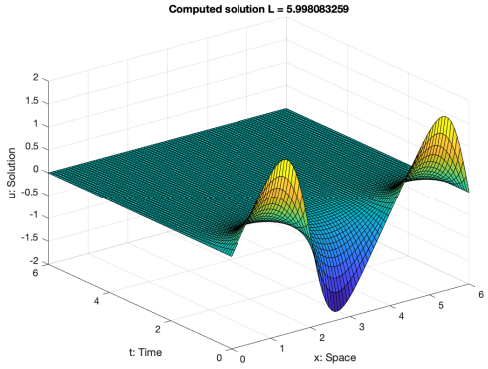


Fig. 6 Burgers equation, $\eta = 0$, fixed $u_0(x)$. The computed solution corresponding to $L_c^1 = 5.998083259$.

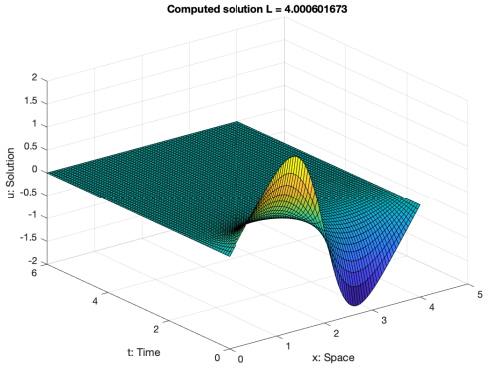


Fig. 7 Burgers equation, $\eta = 0$, fixed $u_0(x)$. The computed solution corresponding to $L_c^2 = 4.000601673$.

Acknowledgements

The author was partially supported by grant PID2020-114976GB-I00 funded by MICIU/AEI/10.13039/501100011033 (Spain).

References

[1] J. Apraiz, J. Cheng, A. Doubova, E. Fernández-Cara and M. Yamamoto. Uniqueness and numerical reconstruction for inverse problems dealing with interval size search. *Inverse Problems and Imaging*, 2022, 16(3): 569–594.

[2] J. Apraiz, A. Doubova, E. Fernández-Cara and M. Yamamoto. Some Inverse Problems for the Burgers Equation and Related Systems. *Communications in Nonlinear Science and Numerical Simulation*, Volume 107, 2022, 106113.

[3] J. Apraiz, A. Doubova, E. Fernández-Cara and M. Yamamoto. Inverse Problems for One-Dimensional Fluid–Solid Interaction Models. *Communications on Applied Mathematics and Computation*, to appear.

[4] M. Bellassoued, M. Yamamoto. Carleman estimates and applications to inverse problems for hyperbolic systems. *Springer Monographs in Mathematics*, Springer, Tokyo, 2017.

[5] A.L. Bukhgeim, J. Cheng, M. Yamamoto. Stability for an inverse boundary problem of determining a part of a boundary. *Inverse Problems*, 1999; 15(4): 1021–1032.

[6] P.P. Carvalho, A. Doubova, E. Fernández-Cara and J. Rocha. Some new results for geometric inverse problems with the method of fundamental solutions. *Inverse Problems in Science and Engineering*, **29** (1), (2021), 131–152.

[7] A. Doubova and E. Fernández-Cara. Some geometric inverse problems for the linear wave equation. *Inverse Problems and Imaging*, **9** (2), (2015), 371–393.

[8] A. Doubova and E. Fernández-Cara. Some geometric inverse problems for the Lamé system with applications in elastography. *Applied Mathematics and Optimization*, (2018), 1–21.

[9] M. Hanke. A taste of inverse problems — basic theory and examples. *Society for Industrial and Applied Mathematics (SIAM)*, Philadelphia, PA, 2017.

[10] A. Hasanov, V.G. Romanov. Introduction to inverse problems for differential equations. *Springer, Cham*, 2017.

[11] F. Hecht. New development in FreeFem++. *Journal of numerical mathematics*, 20.3-4 (2012): 251–266.

[12] X. Huang, O. Imanuvilov and M. Yamamoto. Stability for inverse source problems by Carleman estimates. *Inverse Problems*, **36** (12), (2020), 125006.

[13] MH. Gu, CM Fan, DL Young. The method of fundamental solutions for the multidimensional wave equations. *Journal of Marine Science and Technology*, 2011; 19 (6): 586–595.

[14] A. Karageorghis, D. Lesnic, L. Marin. A survey of applications of the mfs to inverse problems. *Inverse Problems Sci Eng*, 2011; 19: 309–336.

[15] O. Kavian. Lectures on parameter identification. *IRMA Lect Math Theor Phys*, 2003; 4: 125–162.

[16] V. D. Kupradze, M. A. Aleksidze. The Method of Functional Equations for the Approximate Solution of Certain Boundary Value Problems. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 4, (1964), 82–126.

[17] V. Isakov, Inverse Problems for Partial Differential Equations. *Springer, New York*, 2006.

[18] V.G. Romanov. Investigation methods for inverse problems, *Inverse and Ill-posed Problems Series*, VSP, Utrecht, 2002.

[19] M. Yamamoto. Carleman estimates for parabolic equations and applications. *Inverse Problems*, **25** (12), (2009), 123013, 75 pp.

French-German-Spanish Conference on Optimization
Gijón, June 18-21, 2024
(pp. 55-62)

Two results on the control of fluids

Enrique Fernández-Cara
cara@us.es Dept. EDAN and IMUS, University of Sevilla, Spain

Abstract

This talk is devoted to recall and comment two results recently obtained concerning the control of viscous fluids. In general terms, we fix an initial state and we try to find internal or boundary data such that an associated solution vanishes at a prescribed time. Among others, we will consider fluids modeled by coupled systems of the Boussinesq kind. It will be seen that, under some circumstances, the systems are controllable or "quasi-controllable" in an appropriate sense. We will also take a look at some minimal time control problems and will present several theoeretical and numerical results. Also, several open problems will be mentioned.

1. Introduction

Control theory is a scientific discipline that, roughly speaking, try to find out "how can we act on systems". It is multidisciplinary and strongly motivated by real-world applications and, of course, involves mathematics, physics, engineering and other disciplines.

In fact, control problems and their analysis have been considered since ancient times. Among other examples, we can mention

- The irrigation systems (Mesopotamia, since 6,000 B.C.), where the aim was to make reach and stay water regularly in the region.
- The Roman aqueducts (from II Century B.C. to IV Century A.C.), with a similar goal.
- The steam engine (about 1700), an invention that changed the world: the first device able to convert heat into motion in a controlled way.

At present, control theory is applied to problems coming from many different fields:

- From structural mechanics, where control techniques serve to stabilize dangerous and/or undesirable vibrations.
- From the autonomous car driving sector, where very interesting multi-objective problems appear naturally.
- From epidemics and pandemics studies, concerning (for instance) the optimization of vaccination and quarantine strategies.
- From population dynamics, where the objective is to design optimal feeding, tranporting or spreading processes.
- From biomedical sciences, where control is usually oriented to therapy. This applies to cancer, diabetes, alzheimer, etc.

In fluid mechanics, control problems are also very relevant. They are found when one tries to reduce as much as possible turbulence effects, to ensure fluid transportation or design optimal parachutes, wind tunnels and airfcrafts.

In this work, we will be motivated by the control of the Navier-Stokes and the Boussinesq systems. Recall that these PDEs model the behavior of incompressible homogeneous Newtonian fluid respectively insensitive and subject to heat effects.

In the boundary controlled case, they read as follows:

$$\begin{cases} \mathbf{u}_t + (\mathbf{u}\cdot\nabla)\mathbf{u} - \nu\Delta\mathbf{u} + \nabla p = 0, \ \ \nabla\cdot\mathbf{u} = 0, \ \ (\mathbf{x},t) \in \Omega\times(0,T), \\ \mathbf{u} = \mathbf{f}1_\gamma, \ \ (\mathbf{x},t)\in\partial\Omega\times(0,T), \\ + \ \ldots \end{cases}$$

 ISBN 978-84-10135-30-7

and

$$\begin{cases} \mathbf{u}_t + (\mathbf{u}\cdot\nabla)\mathbf{u} - \nu\Delta\mathbf{u} + \nabla p = \theta\mathbf{k}, \ \nabla\cdot\mathbf{u} = 0, \ (\mathbf{x},t)\in\Omega\times(0,T), \\ \theta_t + \mathbf{u}\cdot\nabla\theta - \kappa\Delta\theta = 0, \ (\mathbf{x},t)\in\Omega\times(0,T), \\ \mathbf{u} = \mathbf{f}1_\gamma, \ \theta = \beta 1_\gamma, \ (\mathbf{x},t)\in\partial\Omega\times(0,T), \\ + \ \dots \end{cases}$$

where $\Omega\subset\mathbf{R}^N$ is a bounded open connected domain with Lipschitz-continuous bounndary, $T>0$, $\mathbf{u}$, p and θ respectively denote the velocity field, pressure and temperature of the fluid, $\mathbf{f}$ and β are the controls, γ is a (small) part of $\partial\Omega$, ν and κ are positive constants, $\mathbf{k}$ is a constant vector and the dots contain conditions at $t=0$.

As shown below, the choice of $\mathbf{f}$ and β can be motivated by many different reasons.

2. The null controllability problem for the Boussinesq PDEs

We will first deal with the nonlinear system

$$\begin{cases} \text{Boussinesq for } (\mathbf{u},p,\theta), \ (\mathbf{x},t)\in\Omega\times(0,T), \\ \mathbf{u} = \mathbf{f}1_\gamma, \ \theta = \beta 1_\gamma, \ (\mathbf{x},t)\in\partial\Omega\times(0,T), \\ \mathbf{u}(\cdot,0) = \mathbf{u}_0, \ \theta(\cdot,0) = \theta_0. \end{cases}$$

Let $\mathbf{n}$ stand for the unit outwards normal vector on $\partial\Omega$ and let us intoduce the Hilbert spaces

$$H := \{\mathbf{v}\in L^2(\Omega)^N : \nabla\cdot\mathbf{v} = 0, \ \mathbf{v}\cdot\mathbf{n}|_{\partial\Omega} = 0\}$$

and

$$V := \{\mathbf{v}\in H_0^1(\Omega)^N : \nabla\cdot\mathbf{v} = 0, \ \mathbf{v}|_{\partial\Omega} = 0\},$$

respectively endowed with the norms of $L^2(\Omega)^N$ and $H_0^1(\Omega)^N$.

The problem considered in this section is the following:

For any given $\mathbf{u}_0\in H$ and $\theta_0\in L^2(\Omega)$, find $\mathbf{f}$ and h in appropriate spaces and an associated solution such that

$$\mathbf{u}(\cdot,T) = \mathbf{0} \ \text{ and } \ \theta(\cdot,T) = 0. \tag{2.1}$$

That this problem can be solved was conjectured by J.-L. Lions for the Navier-Stokes PDEs in 1990, see [15]. Since then, many partial (positive) results have been obtained, see among others [1–4, 6, 7, 9, 11–14].

However, the conjecture is open. Even more, whether or not the following approximative version is solvable is also unknown:

For any given $\mathbf{u}_0\in H$ and $\theta_0\in L^2(\Omega)$ and any $\varepsilon>0$, find $\mathbf{f}_\varepsilon$ and h_ε in appropriate spaces and an associated solution such that

$$\|(\mathbf{u},\theta)(\cdot,T)\|\le\varepsilon$$

(here and henceforth, $\|\cdot\|$ stands for the usual L^2 norm in Ω).

We will assume that $N=3$, Ω is a cube and γ is the complement of a face and we will recall a result that proves the conjecture in a different "approximate" sense. More precisely, the following result holds:

Theorem 2.1 *Let $(\mathbf{u}_0,\theta_0)\in H\times L^2(\Omega)$ be given. There exist a family of "ghost" right hand sides $\{(\mathbf{F}_\epsilon,G_\epsilon)\}$ with $(\mathbf{F}_\varepsilon,G_\varepsilon)\to(0,0)$ in an appropriate (large) space as $\varepsilon\to 0$ such that, for any ε there exist controls $(\mathbf{f}_\varepsilon,h_\varepsilon)$ and associated solutions to the nonhomogeneous systems*

$$\begin{cases} \mathbf{u}_t - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = \theta\mathbf{k} + \mathbf{F}_\epsilon, \quad \nabla\cdot\mathbf{u} = 0, \ (\mathbf{x},t)\in\Omega\times(0,T), \\ \theta_t - \kappa\Delta\theta + \mathbf{u}\cdot\nabla\theta = G_\epsilon, \ (\mathbf{x},t)\in\Omega\times(0,T), \\ \mathbf{u} = \mathbf{f}_\varepsilon 1_\gamma, \quad \theta = \beta_\varepsilon 1_\gamma, \ (\mathbf{x},t)\in\partial\Omega\times(0,T), \\ + \ \dots \end{cases}$$

such that (2.1) *holds.*

SKETCH OF THE PROOF:

The complete proof is given in [13] (resp. [7]) in the case of the Navier-Stokes (resp. Boussinesq) system.

For example, let us assume that $\Omega=(-2,2)^3$ and γ is the complement to the face $\{x_1=-2\}$ (see Fig. 1) and let $\varepsilon>0$ be given.

The proof consists of four steps:

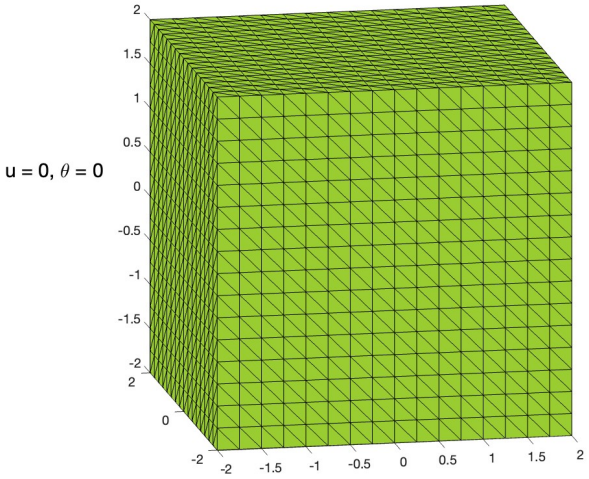


Fig. 1 The control acts on γ, the complement of the left vertical face.

Step 1: Take T_1 close to T and zero controls in $[0, T_1]$ and let the system evolve from $t = 0$ to $t = T_1$.

We can assume that $(\mathbf{u}_1, \theta_1) := (\mathbf{u}, \theta)(\cdot\,, T_1) \in (V \cap H^2(\Omega)^3) \times H^2(\Omega)$. Thus, in this step, the controls and the ghost vanish.

Step 2: Take T_2 in (T_1, T) and a ghost $(\mathbf{F}_\epsilon, G_\epsilon)$ in $[T_1, T_2]$ such that $(\mathbf{u}_2, \theta_2) := (\mathbf{u}, \theta)(\cdot\,, T_2)$ is regular enough and compactly supported in $\overline{\Omega}$.

In this step, the control is zero and it can be assumed that, for some norm $\|(\mathbf{F}_\epsilon, G_\epsilon)\|_*$ is $O(\varepsilon)$ for some appropriate norm $\|\cdot\|_*$.

Step 3: Now, we introduce T_3 in (T_2, T) and a new ghost $(\mathbf{F}_\epsilon, G_\epsilon)$ in $[T_2, T_3]$ such that, at time T_3, the state $(\mathbf{U}, \Theta) := (\mathbf{u}, \theta)(\cdot\,, T_3)$ is of the form $((\phi, 0, \psi), \zeta)(\cdot\,, T_3)$ for functions ϕ, ψ and ζ that only depend on x_2 and t.

This is obviously the main step and, again, we can assume that leads to a small ghost, that is, $\|(\mathbf{F}_\epsilon, G_\epsilon)\|_* = O(\varepsilon)$.

Step 4: Finally, in $[T_3, T]$ we control a 1D heat PDE coupled to a 1D, 2×2 parabolic system.

More precisely, we consider the system

$$\begin{cases} \phi_t - \nu\phi_{x_2,x_2} = 0, \ \ (x_2, t) \in (-2, 2) \times (0, T), \\ \psi_t - \nu\psi_{x_2,x_2} = -\zeta, \quad \zeta_t - \kappa\zeta_{x_2,x_2} = 0, \ \ (x_2, t) \in (-2, 2) \times (0, T), \\ \phi|_{x_2=-2} = \psi|_{x_2=-2} = \zeta|_{x_2=-2} = 0, \ \phi|_{x_2=2} = k(t), \ \psi|_{x_2=2} = \ell(t), \ \zeta|_{x_2=2} = q(t), \ \ t \in (0, T), \\ + \ \ldots \end{cases}$$

where we choose the boundary controls k, ℓ and q such that $((\phi, 0, \psi), \zeta)(\cdot\,, T) = (\mathbf{0}, 0)$.

This can be done in view of the results in [8]. We see that, in this last step, the ghost is zero.

It is clear that, after this construction, the proof is done; the previous argument is illustrated in Fig. 2. □

Some remarks and related open questions are in order:

- The argument in the proof is also applicable to a domain with a flat piece of boundary, see for instance Fig. 3. It would be interesting to deduce an extension to other boundaries, maybe taking into account an appropriate variable change.
- It is possible to improve the result in several directions. Thus, only two scalar controls are needed in Step 4 and this means that Theorem 2.1 remains true with two components of $\mathbf{f}_\varepsilon$ equal to zero; also, the control region can be composed of only three faces; on the other hand, it is found in [5] that a similar result can be deduced in dimension two in a rectangle with ghosts $(\mathbf{F}_\epsilon, G_\epsilon)$ that converge much better to zero, etc.

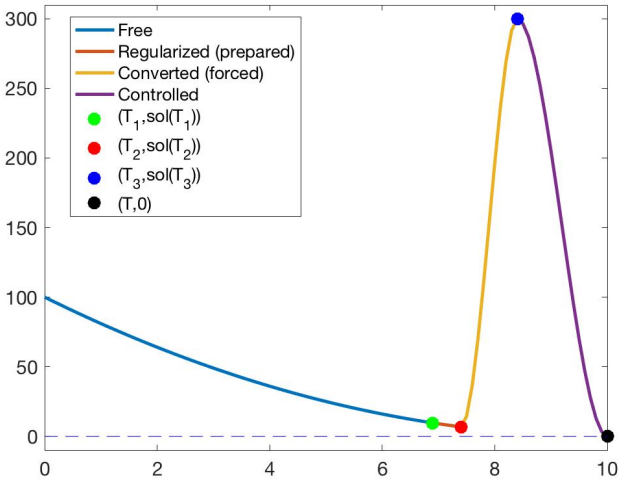


Fig. 2 Main idea: drive $(\mathbf{u}, \theta)$ at $t = T_3$ exactly to $(\mathbf{U}, \Theta)$.

- It is unknown whether the same result can be proved for the so called "full" Boussinesq system, where the heat PDE is
$$\theta_t + \mathbf{u} \cdot \nabla\theta - \kappa\Delta\theta = \nu D\mathbf{u} : \nabla\mathbf{u}.$$
- If we view this result as a preliminary step in the proof of null controllability, it becomes clear that what remains to do is to get an estimate of $(\mathbf{h}_\varepsilon, \beta_\varepsilon)$ somewhere independent of ε. But this is at present unknown.
- On the other hand, it would be interesting to figure out whether this result implies approximate controllability in the sense indicated above. Again, this is an open question.

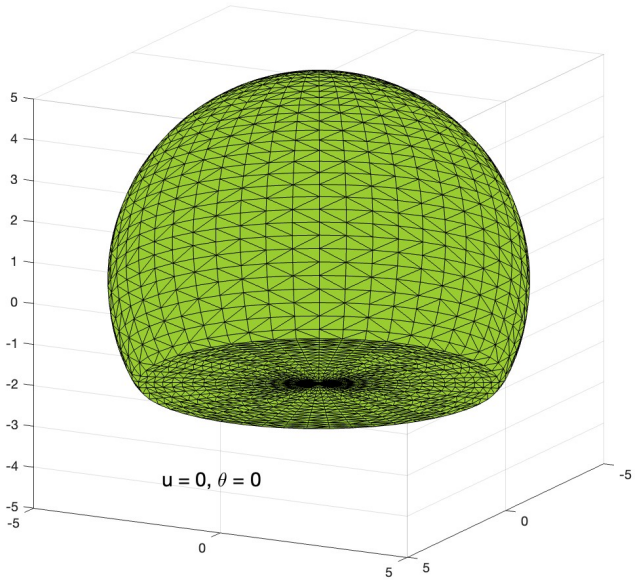


Fig. 3 The control acts on γ, the complement of the flat boundary.

3. A minimal time control problem

In this section, we will be concerned with a control problem where the goal is to find the minimal time needed to drive the system near to a desired state.

For simplicity, we will consider the Navier-Stokes PDEs with controls in the right hand side. Thus, let $\Omega \subset \mathbf{R}^N$ be a bounded connected open set with (for instance) Lipschitz-continuous boundary, let $\omega \subset\subset \Omega$ be a (small) open set and consider the controlled system

$$\begin{cases} \mathbf{u}_t + (\mathbf{u}\cdot\nabla)\mathbf{u} - \nu\Delta\mathbf{u} + \nabla p = \mathbf{f}1_\omega, \ \nabla\cdot\mathbf{u} = 0, \ (\mathbf{x},t) \in \Omega\times(0,T), \\ \mathbf{u} = \mathbf{0}, \ (\mathbf{x},t) \in \partial\Omega\times(0,T), \\ \mathbf{u}(\cdot,0) = \mathbf{u}_0. \end{cases} \tag{3.1}$$

The problem is the following:

Given $\mathbf{u}_0, \mathbf{u}_T \in H$ and $\delta > 0$, find the minimal time $T > 0$ satisfying

$$\|\mathbf{u}(\cdot,T) - \mathbf{u}_T\| \leq \delta, \tag{3.2}$$

where $\mathbf{u}$ is, together with some p, a solution to (3.1) *for some $\mathbf{f}$.*

For general $\mathbf{u}_0$ and $\mathbf{u}_T$ in H, the existence of times T such that (3.2) holds is unknown. But, even if they exist, it is not clear at all that a minimal time can be found.

These considerations justify the following approximated (or penalized) version of the problem:

$$\begin{cases} \text{Minimize } \Phi(T,\mathbf{f}) := \dfrac{T^2}{2} + \dfrac{b}{2}\displaystyle\iint_{\omega\times(0,+\infty)} |\mathbf{f}|^2\,dx\,dt \\ \text{Subject to: } (T,\mathbf{f}) \in \mathcal{H}_{ad}, \end{cases} \tag{3.3}$$

where $b > 0$ and

$$\mathcal{H}_{ad} := \{(T,\mathbf{f}) : T \geq 0, \ \mathbf{f} \in L^2(\omega\times(0,+\infty))^N, \ \|\mathbf{u}(\cdot,T) - \mathbf{u}_T\| = \delta\}.$$

Thus, it can be a good strategy to solve (3.3) for any $b > 0$, then take $b \to 0^+$ and see what happens.

Note that, in several particular situations, the assumption $\mathcal{H}_{ad} \neq \emptyset$ can be asserted; for instance, if $\|\mathbf{u}_T\| < \delta$ this is the case.

In a work in collaboration with I. Marín-Gayte (see [10]), a set of results have been obtained for (3.3):

- First, it is proved that, if $\mathcal{H}_{ad} \neq \emptyset$, there exist optimal couples $(T,\mathbf{f})$.

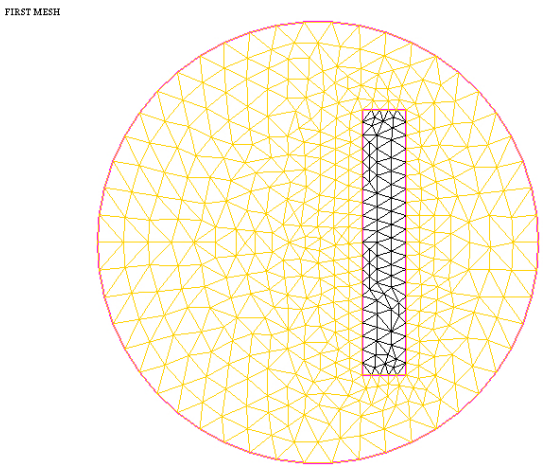


Fig. 4 A numerical test. The domain, the control region and the mesh. Number of points: 1287.

- Then, it is found that, if $(T, \mathbf{f})$ is optimal and an associated $\mathbf{u}$ is regular enough, a suitable optimality system is satisfied. More precisely, the following coupled system must hold for T, $\mathbf{f}$, $\mathbf{u}$, p, a *multiplier* λ and the *adjoint variables* $\mathbf{z}$ and q:

$$\begin{cases} \text{Classical OS for given } \lambda \text{ and } T: \\ \quad \begin{bmatrix} \text{Navier-Stokes for } (\mathbf{u}, p) \text{ and } \mathbf{f} \\ \text{Ajoint system for } (\mathbf{z}, q) \text{ and } \mathbf{u} \\ \ldots \\ \text{Pontryagin for } \mathbf{f} \text{ and } \mathbf{z}\text{: } \mathbf{f} = -\dfrac{1}{\lambda b}\mathbf{z}\Big|_{\omega\times(0,T)} \end{bmatrix} \\ \\ \text{Additional conditions for } \lambda \text{ and } T: \\ \quad \begin{bmatrix} \|\mathbf{u}(\cdot, T) - \mathbf{u}_T\| = \delta \\ T = -\lambda(\mathbf{u}_t(\cdot, T), \mathbf{u}(\cdot, T) - \mathbf{u}_T) \end{bmatrix} \end{cases}$$

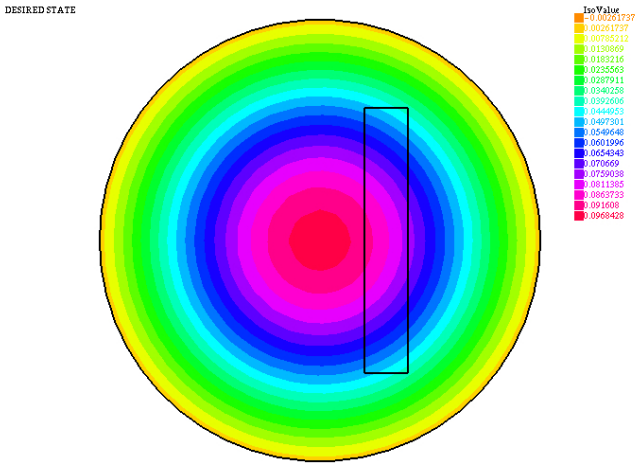


Fig. 5 A numerical test. The target θ_T.

For the proof, the Dubovitsky-Milyutin principle can be used. The argument reads as follows:

- First, we note that, if $(T, \mathbf{f})$ is optimal, there cannot exist descent directions for Φ at the same time *admissible* for the imposed constraints. Consequently, the intersection of the associated cones of directions must be empty.
- By duality, we find that a nontrivial linear combination of dual directions must vanish. After a reformulation and a detailed analysis, we deduce that there exist λ and $(\mathbf{z}, q)$ satisfying the previous optimality system.

- Finally, under the assumption $\mathcal{H}_{ad} \neq \emptyset$, efficient numerical methods are exhibit for the computation of (an approximation to) an optimal $(T, \mathbf{f})$. Among others, the following strategy is proposed:

1. In a first (preliminar) step, for a lot of (λ, T), compute the solution to

$$\text{OC problem} \begin{cases} \text{Navier-Stokes for } (\mathbf{u}, p) \text{ and } \mathbf{f} \\ \text{Ajoint system for } (\mathbf{z}, q) \text{ and } \mathbf{u} \\ + \ \ldots \\ \text{Pontryagin for } \mathbf{f} \text{ and } \mathbf{z}\text{: } \mathbf{f} = -\dfrac{1}{\lambda b}\mathbf{z}\Big|_{\omega\times(0,T)} \end{cases}$$

2. Then, solve the 2×2 system

$$\begin{cases} \|\mathbf{u}(\cdot, T) - \mathbf{u}_T\| = \delta \\ T = -\lambda(\mathbf{u}_t(\cdot, T), \mathbf{u}(\cdot, T) - \mathbf{u}_T) \end{cases}$$

This gives a (candidate to) solution λ_b and T_b.

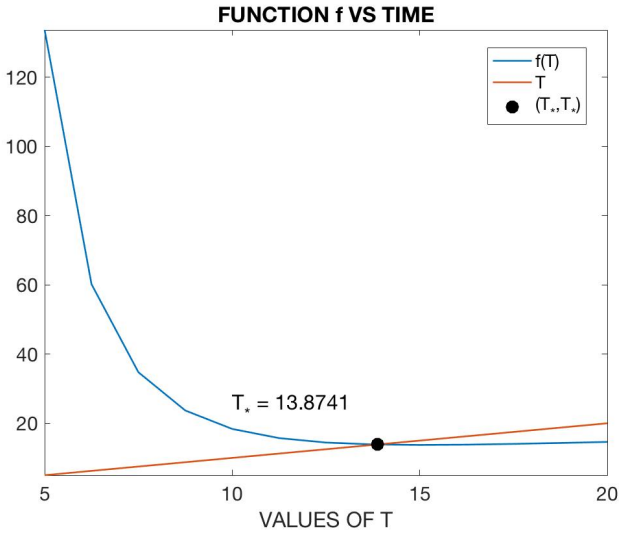


Fig. 6 A numerical test. Solving the equation $T = -\lambda(\theta_t(\cdot,T),\theta(\cdot,T)-\theta_T)$. Computed values: $T = 13.8741$, $\mu = 199.374$, $\lambda = 2.5142\times 10^5$.

In order to illustrate the method, let us give the results of an experiment for a simpler model. Specifically, we consider the problem

$$\begin{cases} \text{Minimize } \Xi(T,f) := \dfrac{T^2}{2} + \dfrac{b}{2}\displaystyle\iint_{\omega\times(0,+\infty)} |f|^2\,dx\,dt, \\ \text{Subject to: } (T,f)\in\mathcal{K}_{ad} \end{cases} \tag{3.4}$$

where

$$\mathcal{K}_{ad} := \{(T,f) : T\geq 0,\ f\in L^2(\omega\times(0,+\infty)),\ \|\theta(\cdot,T)-\theta_T\| = \delta\}$$

and, for each $f\in L^2(\omega\times(0,+\infty))$, we denote by θ the associated solution to

$$\begin{cases} \theta_t - \Delta\theta = f1_\omega, \ \ (\mathbf{x},t)\in\Omega\times(0,T), \\ \theta = 0, \ \ (\mathbf{x},t)\in\partial\Omega\times(0,T), \\ \theta|_{t=0} = 0. \end{cases} \tag{3.5}$$

The domain, the control region and the mesh are depicted in Fig. 4. For the numerical solution of the state systems (3.5) we have used standard implcit Euler finite different approximations on time and $\mathbb{P}_1$ finite element techniques in space.

We have taken

$$b = 100,\ \delta = 0.05 \ \text{ and } \ \theta_T(\mathbf{x}) = R^2 - |\mathbf{x}|^2$$

(see Fig. 5).

The computation of T, μ and $\lambda = 1/(b(2\mu-1))$ is explained in Fig. 6 and the computed final state $\theta(\cdot,T)$ is given in Fig. 7.

We will end this section with several comments and additional questions:

- First, note that we can formulate and solve similar problems for the Boussinesq system, the variable density Navier-Stokes PDEs, boundary controlled models, etc.
- Obviously, the most interesting question is what happens to the solution to (3.3) as $b\to 0^+$? More precisely, when can we ensure uniform estimates of the solutions $(T,\mathbf{f})$ under the assumption $\mathcal{H}_{ad}\neq\emptyset$?
- Finally, observe that an unexplored variant of the considered control problem consists of searching for the minimal time to escape from $\mathbf{u}_0$.

Acknowledgements

The author has been partially supported by grant PID2020-114976GB-I00. This is funded by MICIU/AEI (Spain), with code 10.13039/501100011033.

Fig. 7 A numerical test. The computed $\theta(\cdot, T)$ with $\|\theta(\cdot, T) - \theta_T\| = \delta$.

References

[1] F. W. Chaves-Silva, E. Fernández-Cara, K. Le Balc'h, J. L. F. Machado, and D. A. Souza. Global controllability of the Boussinesq system with Navier-slip-with-friction and Robin boundary conditions. *SIAM J. Control Optim.*, 61(2):484–510, 2023. doi:10.1137/21M1425566.

[2] Jean-Michel Coron. On the controllability of the 2-D incompressible Navier-Stokes equations with the Navier slip boundary conditions. *ESAIM Contrôle Optim. Calc. Var.*, 1:35–75, 1995/96. doi:10.1051/cocv:1996102.

[3] Jean-Michel Coron and Andrei V. Fursikov. Global exact controllability of the 2D Navier-Stokes equations on a manifold without boundary. *Russian J. Math. Phys.*, 4(4):429–448, 1996.

[4] Jean-Michel Coron, Frédéric Marbach, and Franck Sueur. Small-time global exact controllability of the Navier-Stokes equation with Navier slip-with-friction boundary conditions. *J. Eur. Math. Soc. (JEMS)*, 22(5):1625–1673, 2020. doi:10.4171/jems/952.

[5] Jean-Michel Coron, Frédéric Marbach, Franck Sueur, and Ping Zhang. Controllability of the Navier-Stokes equation in a rectangle with a little help of a distributed phantom force. *Ann. PDE*, 5(2):Paper No. 17, 49, 2019. doi:10.1007/s40818-019-0073-4.

[6] E. Fernández-Cara, S. Guerrero, O. Yu. Imanuvilov, and J.-P. Puel. Local exact controllability of the Navier-Stokes system. *J. Math. Pures Appl. (9)*, 83(12):1501–1542, 2004. doi:10.1016/j.matpur.2004.02.010.

[7] Enrique Fernández-Cara, Ivaldo T. De Sousa, and Franciane B. Viera. Remarks concerning the approximate controllability of the 3D Navier-Stokes and Boussinesq systems. *SeMA J.*, 74(3):237–253, 2017. doi:10.1007/s40324-017-0111-7.

[8] Enrique Fernández-Cara, Manuel González-Burgos, and Luz de Teresa. Boundary controllability of parabolic coupled equations. *J. Funct. Anal.*, 259(7):1720–1758, 2010. doi:10.1016/j.jfa.2010.06.003.

[9] Enrique Fernández-Cara, Sergio Guerrero, Oleg Yu. Imanuvilov, and Jean-Pierre Puel. Some controllability results for the N-dimensional Navier-Stokes and Boussinesq systems with $N - 1$ scalar controls. *SIAM J. Control Optim.*, 45(1):146–173, 2006. doi:10.1137/04061965X.

[10] Enrique Fernández-Cara and Irene Marín-Gayte. Analysis and numerical solution of some minimal time control problems. *To appear*, 2025.

[11] A. V. Fursikov and O. Yu. Èmanuilov. Exact controllability of the Navier-Stokes and Boussinesq equations. *Uspekhi Mat. Nauk*, 54(3(327)):93–146, 1999. doi:10.1070/rm1999v054n03ABEH000153.

[12] S. Guerrero. Local exact controllability to the trajectories of the Boussinesq system. *Ann. Inst. H. Poincaré C Anal. Non Linéaire*, 23(1):29–61, 2006. doi:10.1016/j.anihpc.2005.01.002.

[13] Sergio Guerrero, O. Yu. Imanuvilov, and J.-P. Puel. A result concerning the global approximate controllability of the Navier-Stokes system in dimension 3. *J. Math. Pures Appl. (9)*, 98(6):689–709, 2012. doi:10.1016/j.matpur.2012.05.008.

[14] Oleg Yu. Imanuvilov. Remarks on exact controllability for the Navier-Stokes equations. *ESAIM Control Optim. Calc. Var.*, 6:39–72, 2001. doi:10.1051/cocv:2001103.

[15] J.-L. Lions. Remarques sur la controlâbilite approchée. In *Spanish-French Conference on Distributed-Systems Control (Spanish) (Málaga, 1990)*, pages 77–87. Univ. Málaga, Málaga, 1990.

French-German-Spanish Conference on Optimization
Gijón, June 18-21, 2024
(pp. 63-70)

On robust mathematical programs with vanishing constraints with uncertain data

Priyanka Bharati[1], **Vivek Laha**[2]

1. priyankabharati09275@gmail.com Banaras Hindu University, India.
2. laha.vivek333@gmail.com Banaras Hindu University, India. Corresponding author

Abstract

The main objective of this presentation is to explore mathematical programs that incorporate data uncertainty in the vanishing constraints (UMPVC) and to solve them by using a robust optimization framework to deal with the worst-case scenario. To begin with, we derive robust Fritz-John conditions for the UMPVCs and introduce extended no nonzero abnormal multiplier constraint qualification to obtain robust Karush-Kuhn-Tucker conditions. We also identify the robust strong stationary points of the UMPVC and attain sufficient optimality conditions under generalized convexity assumptions. We also identify robust weak stationary points of the UMPVC using a tightened nonlinear programming approach to seek necessary and sufficient robust optimality conditions. The robust version of several constraint qualifications (CQ), like Abadie CQ, Mangasarian-Fromovitz CQ, and linearly independent CQ, are introduced to handle the uncertainties associated with the special structure of the vanishing constraints. Several algorithms are given to apply the results and various examples are presented to illustrate the algorithms.

1. Introduction

A class of nonlinear optimization problems known as mathematical programs with vanishing constraints (MPVC), was initially introduced in [1]. MPVCs are not only rooted in modeling optimal topology design for mechanical structures but also applicable in various technical domains such as mixed-integer nonlinear optimal control problems [12], economic dispatch problems [11], and robot motion planning [13]. However, MPVCs present conceptual and numerical challenges, as standard optimization techniques often struggle due to the combinatorial nature of the vanishing constraints. The difficulties arise from the frequent inability to satisfy standard constraint qualifications, including the Mangasarian-Fromovitz constraint qualification (MFCQ) and the linearly independent constraint qualification (LICQ), at interesting feasible points [8]. Researchers have explored MPVC-tailored constraint qualifications to derive KKT necessary optimality criteria [5, 6]. [2] approached the problem using topological methods in critical point theory, providing insights into stationary points of the MPVCs. The exact penalty theorem, first-order stationary conditions, and second-order stationary conditions have been addressed in [7] and [4]. Results about both weak and strong duality for MPVCs can be found in [9, 19], while [10] discusses Newton-type methods and optimality conditions.

Recently, within the scope of robust optimization, studies have examined MPECs in the presence of uncertainty in the data within the feasible region [15]. Nevertheless, as far as we are aware, no results have yet been found that discuss the optimality conditions for the uncertain MPVC optimization problems, where the vanishing constraint function involves uncertain parameters. Motivated by the aforementioned findings, we focus on examining the KKT optimality conditions for strong and weak stationary points of the uncertain MPVC. Since handling the data uncertainty of the vanishing constraint functions poses trouble, analyzing such an uncertain optimization problem is frequently difficult. To tackle the proposed uncertain optimization problem, we deploy the robust deterministic methodology to investigate robust optimality conditions.

2. Some details

Let $\mathbb{R}^n$ and $\mathbb{R}^n_+$ be the Euclidean space of dimension n and the nonnegative orthant of $\mathbb{R}^n$, respectively. Consider the following MPVC:

$$\begin{cases} \min \quad \mathfrak{f}(x) & & & \\ \text{subject to} \quad x \in \mathcal{F} := \{x \in \mathbb{R}^n \mid & \mathfrak{H}_i(x) \geq 0, & \forall i \in \mathcal{P} := \{1, 2, \dots, p\}, \\ & \mathfrak{G}_i(x)\mathfrak{H}_i(x) \leq 0, & \forall i \in \mathcal{P}\}, \end{cases} \tag{MPVC}$$

where $\mathfrak{f}, \mathfrak{H}_i, \mathfrak{G}_i : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable functions. A point $\tilde{x}$ within the feasible set $\mathcal{F}$ is a *global minimizer* of the MPVC iff

$$\mathfrak{f}(x) \geq \mathfrak{f}(\tilde{x}), \ \forall x \in \mathcal{F}.$$

 ISBN 978-84-10135-30-7

Any point $\tilde{x} \in \mathcal{F}$ is a *local minimizer* of the MPVC iff there exists $\epsilon > 0$ such that

$$\mathfrak{f}(x) \geq \mathfrak{f}(\tilde{x}),\ \forall x \in \mathcal{B}(\tilde{x}, \epsilon) \cap \mathcal{F}.$$

where $\mathcal{B}(\tilde{x}, \epsilon) := \{x \in \mathbb{R}^n : \|x - \tilde{x}\| \leq \epsilon\}$.

Some constraint qualifications that could be satisfied at an optimal point are given as follows:

Definition 2.1 [1] We say that

- the *Abadie constraint qualification (ACQ)* is satisfied at $\tilde{x} \in \mathcal{F}$ iff $\mathcal{T}(\tilde{x}) = \mathcal{L}(\tilde{x})$, where

$$\mathcal{T}(\tilde{x}) := \left\{ d \in \mathbb{R}^n \middle| \exists \{x_n\} \subseteq \mathcal{F}, \exists \{t_n\} \downarrow 0 : x_n \to \tilde{x} \text{ and } \frac{x_n - \tilde{x}}{t_n} \to d \right\}$$

  is the *standard tangent cone* at any point $\tilde{x} \in \mathcal{F}$ and

$$\begin{aligned} \mathcal{L}(\tilde{x}) := \{d \in \mathbb{R}^n : & \nabla \mathfrak{H}(\tilde{x})^T d = 0, && \forall i \in \mathcal{I}_{0+}(\tilde{x}), \\ & \nabla \mathfrak{H}(\tilde{x})^T d \geq 0, && \forall i \in \mathcal{I}_{00}(\tilde{x}) \cup \mathcal{I}_{0-}(\tilde{x}), \\ & \nabla \mathfrak{G}_i(\tilde{x})^T d \leq 0, && \forall i \in \mathcal{I}_{+0}(\tilde{x})\}, \end{aligned}$$

  is the *linearized cone* of the MPVC at point $\tilde{x} \in \mathcal{F}$ with the indices

$$\begin{aligned} &\mathcal{I}_0(\tilde{x}) := \{i \in \mathcal{P} \ : \ \mathfrak{H}_i(\tilde{x}) = 0\}, && \mathcal{I}_+(\tilde{x}) := \{i \in \mathcal{P} \ : \ \mathfrak{H}_i(\tilde{x}) > 0\}, \\ &\mathcal{I}_{00}(\tilde{x}) := \{i \in \mathcal{P} : \mathfrak{H}_i(\tilde{x}) = 0, \mathfrak{G}_i(\tilde{x}) = 0\}, && \mathcal{I}_{0+}(\tilde{x}) := \{i \in \mathcal{P} : \mathfrak{H}_i(\tilde{x}) = 0, \mathfrak{G}_i(\tilde{x}) > 0\}, \\ &\mathcal{I}_{0-}(\tilde{x}) := \{i \in \mathcal{P} : \mathfrak{H}_i(\tilde{x}) = 0, \mathfrak{G}_i(\tilde{x}) < 0\}, && \mathcal{I}_{+0}(\tilde{x}) := \{i \in \mathcal{P} : \mathfrak{H}_i(\tilde{x}) > 0, \mathfrak{G}_i(\tilde{x}) = 0\}, \\ &\mathcal{I}_{+-}(\tilde{x}) := \{i \in \mathcal{P} : \mathfrak{H}_i(\tilde{x}) > 0, \mathfrak{G}_i(\tilde{x}) < 0\}; \end{aligned}$$

- the *modified Abadie constraint qualification (VC-ACQ)* holds at $\tilde{x} \in \mathcal{F}$ iff

$$\mathcal{L}^{VC}(\tilde{x}) \subseteq \mathcal{T}(\tilde{x}),$$

  where

$$\begin{aligned} \mathcal{L}^{VC}(\tilde{x}) := \{d \in \mathbb{R}^n : & \nabla \mathfrak{H}(\tilde{x})^T d = 0, && \forall i \in \mathcal{I}_{0+}(\tilde{x}), \\ & \nabla \mathfrak{H}(\tilde{x})^T d \geq 0, && \forall i \in \mathcal{I}_{00}(\tilde{x}) \cup \mathcal{I}_{0-}(\tilde{x}), \\ & \nabla \mathfrak{G}_i(\tilde{x})^T d \leq 0, && \forall i \in \mathcal{I}_{00}(\tilde{x}) \cup \mathcal{I}_{+0}(\tilde{x})\}. \end{aligned}$$

Theorem 2.2 *[1, Theorem 1] If $\tilde{x} \in \mathcal{F}$ is a local minimum of the MPVC such that*

- *the ACQ is satisfied at $\tilde{x}$, then there exist $\hat{\lambda}_i^{\mathfrak{H}} \in \mathbb{R}, i \in \mathcal{P}, \hat{\lambda}_i^{\mathfrak{G}} \in \mathbb{R}, i \in \mathcal{P}$ such that*

$$\nabla \mathfrak{f}(\tilde{x}) - \sum_{i=1}^{p} \hat{\lambda}_i^{\mathfrak{H}} \nabla \mathfrak{H}_i(\tilde{x}) + \sum_{i=1}^{p} \hat{\lambda}_i^{\mathfrak{G}} \nabla \mathfrak{G}_i(\tilde{x}) = 0, \tag{2.1}$$

  and

$$\begin{cases} \hat{\lambda}_i^{\mathfrak{H}} \text{free}, i \in \mathcal{I}_{0+}(\tilde{x}), & \hat{\lambda}_i^{\mathfrak{H}} \geq 0, i \in \mathcal{I}_{0-}(\tilde{x}) \cup \mathcal{I}_{00}(\tilde{x}), \quad \hat{\lambda}_i^{\mathfrak{H}} = 0, i \in \mathcal{I}_{+0}(\tilde{x}) \cup \mathcal{I}_{+-}(\tilde{x}), \\ \hat{\lambda}_i^{\mathfrak{G}} \geq 0, i \in \mathcal{I}_{+0}(\tilde{x}), & \hat{\lambda}_i^{\mathfrak{G}} = 0, i \in \mathcal{I}_0(\tilde{x}) \cup \mathcal{I}_{+-}(\tilde{x}); \end{cases} \tag{2.2}$$

- *VC-ACQ holds at $\tilde{x}$, then there exist $\hat{\lambda}_i^{\mathfrak{H}} \in \mathbb{R}, i \in \mathcal{P}, \hat{\lambda}_i^{\mathfrak{G}} \in \mathbb{R}, i \in \mathcal{P}$ such that*

$$\nabla \mathfrak{f}(\tilde{x}) - \sum_{i=1}^{p} \hat{\lambda}_i^{\mathfrak{H}} \nabla \mathfrak{H}_i(\tilde{x}) + \sum_{i=1}^{p} \hat{\lambda}_i^{\mathfrak{G}} \nabla \mathfrak{G}_i(\tilde{x}) = 0, \tag{2.3}$$

  and

$$\begin{cases} \hat{\lambda}_i^{\mathfrak{H}} \text{free}, i \in \mathcal{I}_{0+}(\tilde{x}), & \hat{\lambda}_i^{\mathfrak{H}} \geq 0, i \in \mathcal{I}_{0-}(\tilde{x}) \cup \mathcal{I}_{00}(\tilde{x}), \quad \hat{\lambda}_i^{\mathfrak{H}} = 0, i \in \mathcal{I}_{+0}(\tilde{x}) \cup \mathcal{I}_{+-}(\tilde{x}), \\ \hat{\lambda}_i^{\mathfrak{G}} \geq 0, i \in \mathcal{I}_{00}(\tilde{x}) \cup \mathcal{I}_{+0}(\tilde{x}), & \hat{\lambda}_i^{\mathfrak{G}} = 0, i \in \mathcal{I}_{0-}(\tilde{x}) \cup \mathcal{I}_{0+}(\tilde{x}) \cup \mathcal{I}_{+-}(\tilde{x}). \end{cases} \tag{2.4}$$

3. Main Results

In this section, we deal with the MPVC with data uncertainty in the vanishing constraints $\mathfrak{G}_i(x), i \in \mathcal{P}$ due to either measurement errors or insufficient data. The associated uncertain problem is given by:

$$\begin{cases} \min & \mathfrak{f}(x) \\ \text{s.t.} & \mathfrak{H}_i(x) \geq 0, \qquad \forall i \in \mathcal{P}, \\ & \mathfrak{G}_i(x, u_i)\mathfrak{H}_i(x) \leq 0, \quad \forall i \in \mathcal{P}, \end{cases} \tag{UMPVC}$$

where $u_i \in \mathcal{U}_i$ is the uncertain parameter for any convex compact set $\mathcal{U}_i \subseteq \mathbb{R}^m$, the functions $\mathfrak{f}, \mathfrak{H}_i : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable and the functions $\mathfrak{G}_i : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ are continuously differentiable wrt the first component for every $i \in \mathcal{P}$. The robust counterpart of the UMPVC is given as follows:

$$\begin{cases} \min & \mathfrak{f}(x) \\ \text{s.t.} & x \in \Omega := \{x \in \mathbb{R}^n \mid \mathfrak{H}_i(x) \geq 0, \qquad \forall i \in \mathcal{P}, \\ & \qquad\qquad\quad \mathfrak{G}_i(x, u_i)\mathfrak{H}_i(x) \leq 0, \quad \forall u_i \in \mathcal{U}_i, \forall i \in \mathcal{P}\}. \end{cases} \tag{RMPVC}$$

Any point $\tilde{x} \in \Omega$ is a *robust global minimizer* of the UMPVC iff

$$\mathfrak{f}(x) \geq \mathfrak{f}(\tilde{x}), \ \forall x \in \Omega.$$

Any point $\tilde{x} \in \Omega$ is a *robust local minimizer* of the UMPVC iff there exists $\epsilon > 0$ such that

$$\mathfrak{f}(x) \geq \mathfrak{f}(\tilde{x}), \ \forall x \in \mathcal{B}(\tilde{x}, \epsilon) \cap \Omega.$$

It is worth mentioning that a robust local minimizer of the UMPVC is equivalent to a local minimizer of the RMPVC.

Let us define a function $\widehat{\mathfrak{G}}_i : \mathbb{R}^n \to \mathbb{R}$ such that $\widehat{\mathfrak{G}}_i(x) = \sup_{u_i \in \mathcal{U}_i} \mathfrak{G}_i(x, u_i)$ for each $i \in \mathcal{P}$. From [17, Theorem 2.4], we get

$$\{\nabla \widehat{\mathfrak{G}}_i(\tilde{x})\} = \cup_{u_i \in \mathcal{U}_i(\tilde{x})} \{\nabla_x \mathfrak{G}_i(\tilde{x}, u_i)\},$$

where, $\mathcal{U}_i(\tilde{x}) := \{u_i \in \mathcal{U}_i : \mathfrak{G}_i(\tilde{x}, u_i) = \widehat{\mathfrak{G}}_i(\tilde{x})\}$ for every $i \in \mathcal{P}$, Subsequently, the expression for the RMPVC can be reformulated as follows:

$$\begin{cases} \min & \mathfrak{f}(x) \\ \text{s.t.} & \mathfrak{H}_i(x) \geq 0, \qquad \forall i \in \mathcal{P}, \\ & \widehat{\phi_i}(x) = \widehat{\mathfrak{G}}_i(x)\mathfrak{H}_i(x) \leq 0, \qquad \forall i \in \mathcal{P}. \end{cases} \tag{MPVC2}$$

3.1. Robust strong stationary points of the UMPVC

We establish the standard Fritz-John (FJ) type necessary optimality conditions [18] to identify a robust local minimizer of the UMPVC.

Theorem 3.1 ***(Robust FJ conditions for UMPVC)*** *Suppose $\tilde{x} \in \Omega$ is a robust local minimizer of the UMPVC. Assume that $\mathfrak{G}_i(x, \cdot)$ is concave on $\mathcal{U}_i$ for each $x \in \mathbb{R}^n$ and for each $i \in \mathcal{P}$. Then, there exist $\hat{\lambda}^{\mathfrak{f}} \geq 0, \hat{\lambda}_i^{\mathfrak{H}} \in \mathbb{R}, i \in \mathcal{P}, \hat{\lambda}_i^{\mathfrak{G}} \in \mathbb{R}, i \in \mathcal{P}$, not all zero, and $\tilde{u}_i \in \mathcal{U}_i, \ i \in \mathcal{P}$ such that*

$$\hat{\lambda}^{\mathfrak{f}} \nabla \mathfrak{f}(\tilde{x}) - \sum_{i=1}^{p} \hat{\lambda}_i^{\mathfrak{H}} \nabla \mathfrak{H}_i(\tilde{x}) + \sum_{i=1}^{p} \hat{\lambda}_i^{\mathfrak{G}} \nabla_x \mathfrak{G}_i(\tilde{x}, \tilde{u}_i) = 0, \tag{3.1}$$

and

$$\begin{cases} \hat{\lambda}_i^{\mathfrak{H}} \text{free}, i \in \tilde{\mathcal{J}}_{0+}(\tilde{x}, \tilde{u}_i), \hat{\lambda}_i^{\mathfrak{H}} \geq 0, i \in \tilde{\mathcal{J}}_{0-}(\tilde{x}, \tilde{u}_i) \cup \tilde{\mathcal{J}}_{00}(\tilde{x}, \tilde{u}_i), \quad \hat{\lambda}_i^{\mathfrak{H}} = 0, i \in \tilde{\mathcal{J}}_{+0}(\tilde{x}, \tilde{u}_i) \cup \tilde{\mathcal{J}}_{+-}(\tilde{x}, \tilde{u}_i), \\ \hat{\lambda}_i^{\mathfrak{G}} \geq 0, i \in \tilde{\mathcal{J}}_{+0}(\tilde{x}, \tilde{u}_i), \quad \hat{\lambda}_i^{\mathfrak{G}} = 0, i \in \tilde{\mathcal{J}}_0(\tilde{x}, \tilde{u}_i) \cup \tilde{\mathcal{J}}_{+-}(\tilde{x}, \tilde{u}_i) \end{cases} \tag{3.2}$$

where,

$$\begin{aligned} \tilde{\mathcal{J}}_0(\tilde{x}, \tilde{u}_i) &:= \{i \in \mathcal{P} \ : \ \mathfrak{H}_i(\tilde{x}) = 0\}, & \tilde{\mathcal{J}}_{00}(\tilde{x}, \tilde{u}_i) &:= \{i \in \mathcal{P} : \mathfrak{H}_i(\tilde{x}) = 0, \mathfrak{G}_i(\tilde{x}, \tilde{u}_i) = 0\}, \\ \tilde{\mathcal{J}}_{0+}(\tilde{x}, \tilde{u}_i) &:= \{i \in \mathcal{P} : \mathfrak{H}_i(\tilde{x}) = 0, \mathfrak{G}_i(\tilde{x}, \tilde{u}_i) > 0\}, & \tilde{\mathcal{J}}_{0-}(\tilde{x}, \tilde{u}_i) &:= \{i \in \mathcal{P} : \mathfrak{H}_i(\tilde{x}) = 0, \mathfrak{G}_i(\tilde{x}, \tilde{u}_i) < 0\}, \\ \tilde{\mathcal{J}}_{+0}(\tilde{x}, \tilde{u}_i) &:= \{i \in \mathcal{P} : \mathfrak{H}_i(\tilde{x}) > 0, \mathfrak{G}_i(\tilde{x}, \tilde{u}_i) = 0\}, & \tilde{\mathcal{J}}_{+-}(\tilde{x}, \tilde{u}_i) &:= \{i \in \mathcal{P} : \mathfrak{H}_i(\tilde{x}) > 0, \mathfrak{G}_i(\tilde{x}, \tilde{u}_i) < 0\}. \end{aligned} \tag{3.3}$$

The UMPVC fails to satisfy several constraint qualifications because of its nonconvex feasible set. We require an appropriate constraint qualification to establish a robust KKT condition from the robust FJ conditions for the UMPVC. We suggest an extended version of the no non-zero abnormal multiplier constraint qualification (ENNAMCQ) for the UMPVC, which emanates from the NNAMCQ introduced by [21].

Definition 3.2 We say that RMPVC-ENNAMCQ is satisfied at $\tilde{x} \in \Omega$ iff for any $\tilde{u}_i \in \mathcal{U}_i(\tilde{x}), i \in \mathcal{P}$, one has

$$\begin{cases} -\sum_{i\in\tilde{J}_{0+}(\tilde{x},\tilde{u}_i)\cup\tilde{J}_{0-}(\tilde{x},\tilde{u}_i)\cup\tilde{J}_{00}(\tilde{x},\tilde{u}_i)} \hat{\lambda}_i^{\mathfrak{H}}\nabla\mathfrak{H}_i(\tilde{x}) + \sum_{i\in\tilde{J}_{+0}(\tilde{x},\tilde{u}_i)} \hat{\lambda}_i^{\mathfrak{G}}\nabla_x\mathfrak{G}_i(\tilde{x},\tilde{u}_i) = 0, \\ \hat{\lambda}_i^{\mathfrak{H}} \geq 0 \ (i \in \tilde{J}_{0-}(\tilde{x},\tilde{u}_i) \cup \tilde{J}_{00}(\tilde{x},\tilde{u}_i)), \ \hat{\lambda}_i^{\mathfrak{G}} \geq 0 \ (i \in \tilde{J}_{+0}(\tilde{x},\tilde{u}_i)) \\ \Longrightarrow \\ \hat{\lambda}_i^{\mathfrak{H}} = 0 \ (\tilde{J}_{0+}(\tilde{x},\tilde{u}_i) \cup \tilde{J}_{0-}(\tilde{x},\tilde{u}_i) \cup \tilde{J}_{00}(\tilde{x},\tilde{u}_i)), \ \hat{\lambda}_i^{\mathfrak{G}} = 0 \ (i \in \tilde{J}_{+0}(\tilde{x},\tilde{u}_i)). \end{cases} \tag{RMPVC-ENNAMCQ}$$

We can now give a robust KKT necessary optimality condition for the RMPVC problem using the ENNAMCQ constraint qualification and FJ condition.

Theorem 3.3 ***(Robust KKT conditions for the UMPVC)*** *Suppose $\tilde{x} \in \Omega$ is a robust local minimizer of the UMPVC. Assume that $\mathfrak{G}_i(x,\cdot)$ is concave on $\mathcal{U}_i$ for each $x \in \mathbb{R}^n$ and for each $i \in \mathcal{P}$. If RMPVC-ENNAMCQ holds at $\tilde{x}$, then there exist $\hat{\lambda}_i^{\mathfrak{H}} \in \mathbb{R}, i \in \mathcal{P}, \hat{\lambda}_i^{\mathfrak{G}} \in \mathbb{R}, i \in \mathcal{P}$, and $\tilde{u}_i \in \mathcal{U}_i, \ i \in \mathcal{P}$ such that*

$$\nabla\mathfrak{f}(\tilde{x}) - \sum_{i=1}^{p} \hat{\lambda}_i^{\mathfrak{H}}\nabla\mathfrak{H}_i(\tilde{x}) + \sum_{i=1}^{p} \hat{\lambda}_i^{\mathfrak{G}}\nabla_x\mathfrak{G}_i(\tilde{x},\tilde{u}_i) = 0 \tag{3.4}$$

and

$$\begin{cases} \hat{\lambda}_i^{\mathfrak{H}} free, i \in \tilde{J}_{0+}(\tilde{x},\tilde{u}_i), & \hat{\lambda}_i^{\mathfrak{H}} \geq 0, i \in \tilde{J}_{0-}(\tilde{x},\tilde{u}_i) \cup \tilde{J}_{00}(\tilde{x},\tilde{u}_i), \quad \hat{\lambda}_i^{\mathfrak{H}} = 0, i \in \tilde{J}_{+0}(\tilde{x},\tilde{u}_i) \cup \tilde{J}_{+-}(\tilde{x},\tilde{u}_i), \\ \hat{\lambda}_i^{\mathfrak{G}} \geq 0, i \in \tilde{J}_{+0}(\tilde{x},\tilde{u}_i), & \hat{\lambda}_i^{\mathfrak{G}} = 0, i \in \tilde{J}_{0}(\tilde{x},\tilde{u}_i) \cup \tilde{J}_{+-}(\tilde{x},\tilde{u}_i). \end{cases} \tag{3.5}$$

We define strong stationary points for the RMPVC by adopting the KKT conditions to the UMPVC and following the stationary points concept as outlined in [3] for the MPVC.

Definition 3.4 (Robust strong stationary point of the UMPVC) A point $\tilde{x} \in \Omega$ is cosidered as a *robust strong stationary point* of the UMPVC iff there exist $\hat{\lambda}_i^{\mathfrak{H}} \in \mathbb{R}, i \in \mathcal{P}, \hat{\lambda}_i^{\mathfrak{G}} \in \mathbb{R}, i \in \mathcal{P}$ and $\tilde{u}_i \in \mathcal{U}_i, \ i \in \mathcal{P}$ which fulfill the equations (3.4) and (3.5).

Further, we will need the following indexing based on index sets of (3.3), which is dependent on $\tilde{x} \in \Omega$ and for any $\tilde{u}_i \in \mathcal{U}_i$:

$$\begin{aligned} \mathcal{I}_{0+}^{+} &:= \{i \in \tilde{J}_{0+}(\tilde{x},\tilde{u}_i) \ : \hat{\lambda}_i^{\mathfrak{H}} > 0\}, \quad \mathcal{I}_{0+}^{-} := \{i \in \tilde{J}_{0+}(\tilde{x},\tilde{u}_i) \ : \hat{\lambda}_i^{\mathfrak{H}} < 0\}, \\ \mathcal{I}_{00}^{+} &:= \{i \in \tilde{J}_{00}(\tilde{x},\tilde{u}_i) \ : \hat{\lambda}_i^{\mathfrak{H}} > 0\}, \quad \mathcal{I}_{0-}^{+} := \{i \in \tilde{J}_{0-}(\tilde{x},\tilde{u}_i) \ : \hat{\lambda}_i^{\mathfrak{H}} > 0\}, \\ \mathcal{I}_{+0}^{+} &:= \{i \in \tilde{J}_{+0}(\tilde{x},\tilde{u}_i) \ : \hat{\lambda}_i^{\mathfrak{G}} > 0\}. \end{aligned} \tag{3.6}$$

Theorem 3.5 ***(Robust sufficient optimality condition)*** *Suppose that $\tilde{x} \in \Omega$ is a robust strong stationary point of the UMPVC, i.e., there exist $\hat{\lambda}_i^{\mathfrak{H}} \in \mathbb{R}, i \in \mathcal{P}, \hat{\lambda}_i^{\mathfrak{G}} \in \mathbb{R}, i \in \mathcal{P}$, and $\tilde{u}_i \in \mathcal{U}_i, \ i \in \mathcal{P}$ such that (3.4) and (3.5) are satisfied. If $\mathfrak{f}$ is pseudoconvex at $\tilde{x}$ and $\mathfrak{G}_i(\cdot,\tilde{u}_i)(i \in \mathcal{I}_{+0}^{+}), \ -\mathfrak{H}_i(i \in \mathcal{I}_{0+}^{+} \cup \mathcal{I}_{00}^{+} \cup \mathcal{I}_{0-}^{+}), \ \mathfrak{H}_i(i \in \mathcal{I}_{0+}^{-})$ are quasiconvex at $\tilde{x}$ over Ω, then*

(a) $\tilde{x}$ is a robust global minimizer of the UMPVC whenever $\mathcal{I}_{0+}^{-} \cup \mathcal{I}_{+0}^{+} = \emptyset$;

(b) $\tilde{x}$ is a robust local minimizer of the UMPVC.

3.2. Robust weak stationary points of the UMPVC

In this segment, we try to find out optimality conditions to identify robust weak stationary points of the UMPVC. For any $\tilde{x} \in \Omega$, a tightened nonlinear programming problem associated with the MPVC2 is given as follows:

$$\begin{cases} \min & \mathfrak{f}(x) & \\ \text{s.t.} & \mathfrak{H}_i(x) = 0, & \forall i \in \hat{J}_{0+}(\tilde{x}) \cup \hat{J}_{00}(\tilde{x}), \\ & \mathfrak{H}_i(x) \geq 0, & \forall i \in \hat{J}_{0-}(\tilde{x}) \cup \hat{J}_{+}(\tilde{x}), \\ & \widehat{\mathfrak{G}}_i(x) \leq 0, & \forall i \in \mathcal{P}. \end{cases} \tag{RTNLP($\tilde{x}$)}$$

It is straightforward to observe that the feasible set of the RTNLP($\tilde{x}$) is within the feasible set of the MPVC2. Therefore, if $\tilde{x}$ is a local minimum of the RMPVC, it will also be a local minimum of the RTNLP($\tilde{x}$).We establish a robust FJ type necessary optimality condition to identify a robust local minimizer of the UMPVC by using the standard FJ conditions given in [18] for the RTNLP($\tilde{x}$).

Theorem 3.6 ***(Robust FJ conditions for UMPVC)*** *Suppose $\tilde{x} \in \Omega$ is a robust local minimizer of the UMPVC. Assume that $\mathfrak{G}_i(x,\cdot)$ is concave on $\mathcal{U}_i$ for each $x \in \mathbb{R}^n$ and for each $i \in \mathcal{P}$. Then, there exist $\hat{\lambda}^{\mathfrak{f}} \geq 0, \hat{\lambda}_i^{\mathfrak{H}} \in \mathbb{R}, i \in \mathcal{P}, \hat{\lambda}_i^{\mathfrak{G}} \in \mathbb{R}, i \in \mathcal{P}$, not all zero, and $\tilde{u}_i \in \mathcal{U}_i,\ i \in \mathcal{P}$ such that*

$$\hat{\lambda}^{\mathfrak{f}} \nabla \mathfrak{f}(\tilde{x}) - \sum_{i=1}^{p} \hat{\lambda}_i^{\mathfrak{H}} \nabla \mathfrak{H}_i(\tilde{x}) + \sum_{i=1}^{p} \hat{\lambda}_i^{\mathfrak{G}} \nabla_x \mathfrak{G}_i(\tilde{x}, \tilde{u}_i) = 0, \tag{3.7}$$

and

$$\begin{cases} \hat{\lambda}_i^{\mathfrak{H}} \text{ free}, i \in \tilde{\mathcal{J}}_{0+}(\tilde{x},\tilde{u}_i) \cup \tilde{\mathcal{J}}_{00}(\tilde{x},\tilde{u}_i), & \hat{\lambda}_i^{\mathfrak{H}} \geq 0, i \in \tilde{\mathcal{J}}_{0-}(\tilde{x},\tilde{u}_i), \quad \hat{\lambda}_i^{\mathfrak{H}} = 0, i \in \tilde{\mathcal{J}}_{+}(\tilde{x},\tilde{u}_i), \\ \hat{\lambda}_i^{\mathfrak{G}} \geq 0, i \in \tilde{\mathcal{J}}_{+0}(\tilde{x},\tilde{u}_i) \cup \tilde{\mathcal{J}}_{00}(\tilde{x},\tilde{u}_i), & \hat{\lambda}_i^{\mathfrak{G}} = 0, i \in \tilde{\mathcal{J}}_{0-}(\tilde{x},\tilde{u}_i) \cup \tilde{\mathcal{J}}_{+-}(\tilde{x},\tilde{u}_i) \cup \tilde{\mathcal{J}}_{0+}(\tilde{x},\tilde{u}_i). \end{cases} \tag{3.8}$$

Using RTNLP($\tilde{x}$), we can provide the ENNAMCQ for the RMPVC at $\tilde{x}$.

Definition 3.7 We say that RTNLP-ENNAMCQ is satisfied at $\tilde{x} \in \Omega$ iff for any $\tilde{u}_i \in \mathcal{U}_i(\tilde{x}), i \in \mathcal{P}$, one has

$$\begin{cases} -\sum_{i \in \tilde{\mathcal{J}}_{0+}(\tilde{x},\tilde{u}_i) \cup \tilde{\mathcal{J}}_{00}(\tilde{x},\tilde{u}_i) \cup \tilde{\mathcal{J}}_{0-}(\tilde{x},\tilde{u}_i)} \hat{\lambda}_i^{\mathfrak{H}} \nabla \mathfrak{H}_i(\tilde{x}) \\ +\sum_{i \in \tilde{\mathcal{J}}_{+0}(\tilde{x},\tilde{u}_i) \cup \tilde{\mathcal{J}}_{00}(\tilde{x},\tilde{u}_i)} \hat{\lambda}_i^{\mathfrak{G}} \nabla_x \mathfrak{G}_i(\tilde{x},\tilde{u}_i) = 0, \\ \hat{\lambda}_i^{\mathfrak{H}} \geq 0\ (i \in \tilde{\mathcal{J}}_{0-}(\tilde{x},\tilde{u}_i)),\ \hat{\lambda}_i^{\mathfrak{G}} \geq 0\ (i \in \tilde{\mathcal{J}}_{+0}(\tilde{x},\tilde{u}_i) \cup \tilde{\mathcal{J}}_{00}(\tilde{x},\tilde{u}_i)) \\ \Longrightarrow \\ \hat{\lambda}_i^{\mathfrak{H}} = 0\ (i \in \tilde{\mathcal{J}}_{0+}(\tilde{x},\tilde{u}_i) \cup \tilde{\mathcal{J}}_{00}(\tilde{x},\tilde{u}_i) \cup \tilde{\mathcal{J}}_{0-}(\tilde{x},\tilde{u}_i)),\ \hat{\lambda}_i^{\mathfrak{G}} = 0\ (i \in \tilde{\mathcal{J}}_{+0}(\tilde{x},\tilde{u}_i) \cup \tilde{\mathcal{J}}_{00}(\tilde{x},\tilde{u}_i)). \end{cases}$$

(RTNLP-ENNAMCQ)

We can now give the robust KKT necessary optimality condition for the UMPVC using Theorem 3.6 and constraint qualification RTNLP-ENNAMCQ.

Theorem 3.8 ***(Robust KKT optimality conditions for UMPVC)*** *Suppose $\tilde{x} \in \Omega$ is a robust local minimizer of the UMPVC. Assume that $\mathfrak{G}_i(x,\cdot)$ is concave on $\mathcal{U}_i$ for each $x \in \mathbb{R}^n$ and for each $i \in \mathcal{P}$. If RTNLP-ENNAMCQ holds at $\tilde{x}$, then there exist $\hat{\lambda}_i^{\mathfrak{H}} \in \mathbb{R}, i \in \mathcal{P}, \hat{\lambda}_i^{\mathfrak{G}} \in \mathbb{R}, i \in \mathcal{P}$, and $\tilde{u}_i \in \mathcal{U}_i,\ i \in \mathcal{P}$ such that*

$$\nabla \mathfrak{f}(\tilde{x}) - \sum_{i=1}^{p} \hat{\lambda}_i^{\mathfrak{H}} \nabla \mathfrak{H}_i(\tilde{x}) + \sum_{i=1}^{p} \hat{\lambda}_i^{\mathfrak{G}} \nabla_x \mathfrak{G}_i(\tilde{x}, \tilde{u}_i) = 0 \tag{3.9}$$

and

$$\begin{cases} \hat{\lambda}_i^{\mathfrak{H}} \text{ free}, i \in \tilde{\mathcal{J}}_{0+}(\tilde{x},\tilde{u}_i) \cup \tilde{\mathcal{J}}_{00}(\tilde{x},\tilde{u}_i), & \hat{\lambda}_i^{\mathfrak{H}} \geq 0, i \in \tilde{\mathcal{J}}_{0-}(\tilde{x},\tilde{u}_i), \quad \hat{\lambda}_i^{\mathfrak{H}} = 0, i \in \tilde{\mathcal{J}}_{+}(\tilde{x},\tilde{u}_i), \\ \hat{\lambda}_i^{\mathfrak{G}} \geq 0, i \in \tilde{\mathcal{J}}_{+0}(\tilde{x},\tilde{u}_i) \cup \tilde{\mathcal{J}}_{00}(\tilde{x},\tilde{u}_i), & \hat{\lambda}_i^{\mathfrak{G}} = 0, i \in \tilde{\mathcal{J}}_{0-}(\tilde{x},\tilde{u}_i) \cup \tilde{\mathcal{J}}_{+-}(\tilde{x},\tilde{u}_i) \cup \tilde{\mathcal{J}}_{0+}(\tilde{x},\tilde{u}_i). \end{cases} \tag{3.10}$$

Now we define weak stationary points for the RMPVC by following the above KKT conditions of the RMPVC.

Definition 3.9 *(Robust weak stationary points for the UMPVC)* Any point $\tilde{x} \in \Omega$ is said to be a *robust weak stationary point* for the UMPVC iff there exist $\hat{\lambda}_i^{\mathfrak{H}} \in \mathbb{R}, i \in \mathcal{P}, \hat{\lambda}_i^{\mathfrak{G}} \in \mathbb{R}, i \in \mathcal{P}$ and $\tilde{u}_i \in \mathcal{U}_i,\ i \in \mathcal{P}$ which satisfy equations (3.9) and (3.10).

Alongside the index sets specified in equation (3.3), we require additional indexing based on the $\tilde{x} \in \Omega$.

$$\begin{aligned} \mathcal{I}_{0+}^{+} &:= \{i \in \tilde{\mathcal{J}}_{0+}(\tilde{x},\tilde{u}_i)\ : \hat{\lambda}_i^{\mathfrak{H}} > 0\}; & \mathcal{I}_{0+}^{-} &:= \{i \in \tilde{\mathcal{J}}_{0+}(\tilde{x},\tilde{u}_i)\ : \hat{\lambda}_i^{\mathfrak{H}} < 0\}; \\ \mathcal{I}_{0-}^{+} &:= \{i \in \tilde{\mathcal{J}}_{0-}(\tilde{x},\tilde{u}_i)\ : \hat{\lambda}_i^{\mathfrak{H}} > 0\}; & \mathcal{I}_{00}^{-} &:= \{i \in \tilde{\mathcal{J}}_{00}(\tilde{x},\tilde{u}_i)\ : \hat{\lambda}_i^{\mathfrak{H}} < 0\}; \\ \mathcal{I}_{00}^{+} &:= \{i \in \tilde{\mathcal{J}}_{00}(\tilde{x},\tilde{u}_i)\ : \hat{\lambda}_i^{\mathfrak{H}} > 0\}; & \mathcal{I}_{+0}^{+} &:= \{i \in \tilde{\mathcal{J}}_{+0}(\tilde{x},\tilde{u}_i)\ : \hat{\lambda}_i^{\mathfrak{H}} > 0\}; \\ \mathcal{I}_{00}^{*+} &:= \{i \in \tilde{\mathcal{J}}_{00}(\tilde{x},\tilde{u}_i)\ : \hat{\lambda}_i^{\mathfrak{G}} > 0\}. \end{aligned} \tag{3.11}$$

Theorem 3.10 ***(Robust sufficient optimality conditions using weak stationarity)*** *Let $\tilde{x} \in \Omega$ be a robust weak stationary point for the UMPVC. If $\dagger$ is pseudoconvex at $\tilde{x}$ and $\mathfrak{G}_i(\cdot, \tilde{u}_i)(i \in \mathcal{I}_{+0}^{+} \cup \mathcal{I}_{00}^{*+})$, $-\mathfrak{H}_i(i \in \mathcal{I}_{0+}^{+} \cup \mathcal{I}_{0-}^{+} \cup \mathcal{I}_{00}^{+})$, $\mathfrak{H}_i(i \in \mathcal{I}_{0+}^{-} \cup \mathcal{I}_{00}^{-})$ are quasiconvex at $\tilde{x}$ over Ω, then*

(a) $\tilde{x}$ is a global robust minimizer of the UMPVC whenever $\mathcal{I}_{0+}^{-} \cup \mathcal{I}_{00}^{-} \cup \mathcal{I}_{+0}^{+} \cup \mathcal{I}_{00}^{*+} = \emptyset$;

(b) $\tilde{x}$ is a local robust minimizer of the UMPVC whenever $\mathcal{I}_{00}^{*+} \cup \mathcal{I}_{00}^{-} = \emptyset$.

3.3. Robust constraint qualifications for the UMPVC

This section must find the standard KKT conditions of a strong stationary point for the UMPVC under some constraint qualifications. Furthermore, we will give the relation between those constraint qualifications.

The following lemma gives the standard linearized cone of the UMPVC at a robust local minimizer.

Lemma 3.11 *Suppose $\tilde{x} \in \Omega$ is a robust local minimizer of the UMPVC. Then, the robust linearized cone of the UMPVC at $\tilde{x}$ is given by*

$$\begin{aligned}\mathcal{L}_{RMPVC}(\tilde{x}) := \{d \in \mathbb{R}^n : &\nabla\mathfrak{H}(\tilde{x})^T d = 0, && \forall i \in \tilde{\mathcal{J}}_{0+}(\tilde{x}, \tilde{u}_i),\\ &\nabla\mathfrak{H}(\tilde{x})^T d \geq 0, && \forall i \in \tilde{\mathcal{J}}_{00}(\tilde{x}, \tilde{u}_i) \cup \tilde{\mathcal{J}}_{0-}(\tilde{x}, \tilde{u}_i),\\ &\nabla_x\mathfrak{G}_i(\tilde{x}, \tilde{u}_i)^T d \leq 0, && \forall i \in \tilde{\mathcal{J}}_{+0}(\tilde{x}, \tilde{u}_i)\},\end{aligned}$$

for some $\tilde{u}_i \in \mathcal{U}_i(\tilde{x}), i \in \mathcal{P}$.

Based on the robust linearized cone of the UMPVC, we give an extended version of the Abadie constraint qualification, denoted by EACQ, for the UMPVC.

Definition 3.12 ***(EACQ for the RMPVC)*** Let $\tilde{x}$ be a robust local minimizer of the UMPVC. Then, the EACQ holds at $\tilde{x}$ iff $\mathcal{T}_{RMPVC}(\tilde{x}) = \mathcal{L}_{RMPVC}(\tilde{x})$.

The following theorem gives the standard KKT conditions for a robust local minimizer of the UMPVC when EACQ is satisfied.

Theorem 3.13 *Suppose $\tilde{x}$ is a robust local minimizer of the UMPVC and EACQ satisfied at $\tilde{x}$. Then, there exist $\hat{\lambda}_i^{\mathfrak{H}} \in \mathbb{R}, i \in \mathcal{P}, \hat{\lambda}_i^{\mathfrak{G}} \in \mathbb{R}, i \in \mathcal{P}$ and $\tilde{u}_i \in \mathcal{U}_i,\ i \in \mathcal{P}$ such that*

$$\nabla\dagger(\tilde{x}) - \sum_{i=1}^{p} \hat{\lambda}_i^{\mathfrak{H}} \nabla\mathfrak{H}_i(\tilde{x}) + \sum_{i=1}^{p} \hat{\lambda}_i^{\mathfrak{G}} \nabla_x\mathfrak{G}_i(\tilde{x}, \tilde{u}_i) = 0, \tag{3.12}$$

and

$$\begin{cases}\hat{\lambda}_i^{\mathfrak{H}} free, i \in \tilde{\mathcal{J}}_{0+}(\tilde{x}, \tilde{u}_i), & \hat{\lambda}_i^{\mathfrak{H}} \geq 0, i \in \tilde{\mathcal{J}}_{0-}(\tilde{x}, \tilde{u}_i) \cup \tilde{\mathcal{J}}_{00}(\tilde{x}, \tilde{u}_i), \quad \hat{\lambda}_i^{\mathfrak{H}} = 0, i \in \tilde{\mathcal{J}}_{+0}(\tilde{x}, \tilde{u}_i) \cup \tilde{\mathcal{J}}_{+-}(\tilde{x}, \tilde{u}_i),\\ \hat{\lambda}_i^{\mathfrak{G}} \geq 0, i \in \tilde{\mathcal{J}}_{+0}(\tilde{x}, \tilde{u}_i), & \hat{\lambda}_i^{\mathfrak{G}} = 0, i \in \tilde{\mathcal{J}}_{0}(\tilde{x}, \tilde{u}_i) \cup \tilde{\mathcal{J}}_{+-}(\tilde{x}, \tilde{u}_i).\end{cases} \tag{3.13}$$

We give extended versions of VC-MFCQ, VC-LICQ and VC-ACQ provided by [1] for the UMPVC.

Definition 3.14 Let $\tilde{x} \in \Omega$ be a local robust minimizer of the UMPVC. Then,

(a) VC-EMFCQ is satisfied at $\tilde{x}$ iff the gradients

$$\nabla\mathfrak{H}_i(\tilde{x}), i \in \tilde{\mathcal{J}}_{00}(\tilde{x}, \tilde{u}_i) \cup \tilde{\mathcal{J}}_{0+}(\tilde{x}, \tilde{u}_i)$$

are linearly independent, and there exists a vector $\tilde{d}$ such that

$$\begin{cases}\nabla\mathfrak{H}_i(\tilde{x})^T \tilde{d} = 0, & \forall i \in \tilde{\mathcal{J}}_{00}(\tilde{x}, \tilde{u}_i) \cup \tilde{\mathcal{J}}_{0+}(\tilde{x}, \tilde{u}_i),\\ \nabla\mathfrak{H}_i(\tilde{x})^T \tilde{d} > 0, & \forall i \in \tilde{\mathcal{J}}_{0-}(\tilde{x}, \tilde{u}_i),\\ \nabla_x\mathfrak{G}_i(\tilde{x}, \tilde{u}_i)^T \tilde{d} < 0, & \forall i \in \tilde{\mathcal{J}}_{+0}(\tilde{x}, \tilde{u}_i).\end{cases} \tag{3.14}$$

is satisfied for some $\tilde{u}_i \in \mathcal{U}_i(\tilde{x})$;

(b) VC-ELICQ is satisfied at $\tilde{x}$ iff the gradients

$$\begin{matrix} \nabla_x \mathfrak{G}_i(\tilde{x}, \tilde{u}_i), & i \in \tilde{\mathcal{J}}_{+0}(\tilde{x}, \tilde{u}_i), \\ \nabla \mathfrak{H}_i(\tilde{x}), & i \in \tilde{\mathcal{J}}_0(\tilde{x}, \tilde{u}_i) \end{matrix} \tag{3.15}$$

are linearly independent for any $\tilde{u}_i \in \mathcal{U}_i(\tilde{x})$. We can identify that, if VC-ELICQ holds at $\tilde{x}$, then VC-EMFCQ is also satisfied at $\tilde{x}$;

Now, we define a modified EACQ which is weaker than the EACQ.

Definition 3.15 Let $\tilde{x}$ be the robust local minimizer of the UMPVC. Then, the *modified extended Abadie constraint qualification (VC-EACQ)* holds at $\tilde{x}$ iff

$$\mathcal{L}^{VC}_{RMPVC}(\tilde{x}) \subseteq \mathcal{T}_{RMPVC}(\tilde{x}),$$

where

$$\begin{aligned} \mathcal{L}^{VC}_{RMPVC}(\tilde{x}) := \{d \in \mathbb{R}^n : & \nabla \mathfrak{H}(\tilde{x})^T d = 0, && \forall i \in \tilde{\mathcal{J}}_{0+}(\tilde{x}, \tilde{u}_i), \\ & \nabla \mathfrak{H}(\tilde{x})^T d \geq 0, && \forall i \in \tilde{\mathcal{J}}_{00}(\tilde{x}, \tilde{u}_i) \cup \tilde{\mathcal{J}}_{0-}(\tilde{x}, \tilde{u}_i), \\ & \nabla_x \mathfrak{G}_i(\tilde{x}, \tilde{u}_i)^T d \leq 0, && \forall i \in \tilde{\mathcal{J}}_{00}(\tilde{x}, \tilde{u}_i) \cup \tilde{\mathcal{J}}_{+0}(\tilde{x}, \tilde{u}_i)\} \end{aligned}$$

for some $\tilde{u}_i \in \mathcal{U}_i(\tilde{x}), i \in \mathcal{P}$.

Relation between different constraint qualifications in a flow chart

$$\begin{array}{ccccc} & & & & \boxed{\text{EACQ}} \\ & & & & \Downarrow \\ \boxed{\text{VC-ELICQ}} & \Rightarrow & \boxed{\text{VC-EMFCQ}} & \Rightarrow & \boxed{\text{VC-EACQ}} \end{array}$$

We have the following robust KKT condition under the assumption of VC-EACQ.

Theorem 3.16 *(Robust KKT conditions under VC-EACQ) Let $\tilde{x}$ be a robust local minimizer of the UMPVC such that VC-EACQ holds at $\tilde{x}$. Then, there exist $\hat{\lambda}^{\mathfrak{H}}_i \in \mathbb{R}, i \in \mathcal{P}, \hat{\lambda}^{\mathfrak{G}}_i \in \mathbb{R}, i \in \mathcal{P}$, and $\tilde{u}_i \in \mathcal{U}_i,\ i \in \mathcal{P}$ such that*

$$\nabla \mathfrak{f}(\tilde{x}) - \sum_{i=1}^{p} \hat{\lambda}^{\mathfrak{H}}_i \nabla \mathfrak{H}_i(\tilde{x}) + \sum_{i=1}^{p} \hat{\lambda}^{\mathfrak{G}}_i \nabla_x \mathfrak{G}_i(\tilde{x}, \tilde{u}_i) = 0, \tag{3.16}$$

and

$$\begin{cases} \hat{\lambda}^{\mathfrak{H}}_i \text{free}, i \in \tilde{\mathcal{J}}_{0+}(\tilde{x}, \tilde{u}_i), & \hat{\lambda}^{\mathfrak{H}}_i \geq 0, i \in \tilde{\mathcal{J}}_{0-}(\tilde{x}, \tilde{u}_i) \cup \tilde{\mathcal{J}}_{00}(\tilde{x}, \tilde{u}_i), & \hat{\lambda}^{\mathfrak{H}}_i = 0, i \in \tilde{\mathcal{J}}_{+}(\tilde{x}, \tilde{u}_i), \\ \hat{\lambda}^{\mathfrak{G}}_i \geq 0, i \in \tilde{\mathcal{J}}_{00}(\tilde{x}, \tilde{u}_i) \cup \tilde{\mathcal{J}}_{+0}(\tilde{x}, \tilde{u}_i), & \hat{\lambda}^{\mathfrak{G}}_i = 0, i \in \tilde{\mathcal{J}}_{0+}(\tilde{x}, \tilde{u}_i) \cup \tilde{\mathcal{J}}_{0-}(\tilde{x}, \tilde{u}_i) \cup \tilde{\mathcal{J}}_{+-}(\tilde{x}, \tilde{u}_i). \end{cases} \tag{3.17}$$

4. Conclusion

We have studied mathematical programs with vanishing constraints with uncertainty in the feasible region denoted by UMPVC and used the robust optimization approach to deal with the worst case scenario. First, we use the Fritz-John approach to identify robust strong stationary points of the UMPVC and introduce RMPVC-ENNAMCQ to derive the robust KKT necessary and sufficient optimality conditions. Further, we utilize a tightened nonlinear programming approach to determine robust weak stationary points of the UMPVC. Necessary and sufficient conditions are deived under RTNLP-ENNAMCQ and generalized convexity assumptions. Several other constraint qualifications, like EACQ, VC-EACQ, VC-EMFCQ, VC-ELICQ, are introduced to deal with the uncertainty of the feasible region and relationships among them are given. The results are illustrated with suitable algorithms and examples. Further, the results can be extended for multiobjective optimization [20], minimax programming [14] and nonsmooth problems [16] with vanishing constraints.

References

[1] Wolfgang Achtziger and Christian Kanzow. Mathematical programs with vanishing constraints: optimality conditions and constraint qualifications. Mathematical Programming, 114:69–99, 2008. doi:10.1007/s10107-006-0083-3.

[2] Dominik Dorsch, Vladimir Shikhman, and Oliver Stein. Mathematical programs with vanishing constraints: critical point theory. Journal of Global Optimization, 52:591–605, 2012. doi:10.1007/s10898-011-9805-z.

[3] Tim Hoheisel. *Mathematical programs with vanishing constraints*. PhD thesis, Universität Würzburg, 2009.

[4] Tim Hoheisel and Christian Kanzow. First-and second-order optimality conditions for mathematical programs with vanishing constraints. Applications of Mathematics, 52(6):495–514, 2007. doi:10.1007/s10492-007-0029-y.

[5] Tim Hoheisel and Christian Kanzow. Stationary conditions for mathematical programs with vanishing constraints using weak constraint qualifications. Journal of Mathematical Analysis and Applications, 337(1):292–310, 2008. doi:10.1016/j.jmaa.2007.03.087.

[6] Tim Hoheisel and Christian Kanzow. On the abadie and guignard constraint qualifications for mathematical programmes with vanishing constraints. Optimization, 58(4):431–448, 2009. doi:10.1080/02331930701763405.

[7] Tim Hoheisel, Christian Kanzow, and Jiří V Outrata. Exact penalty results for mathematical programs with vanishing constraints. Nonlinear Analysis: Theory, Methods & Applications, 72(5):2514–2526, 2010. doi:10.1016/j.na.2009.10.047.

[8] Tim Hoheisel, Blanca Pablos, Aram Pooladian, Alexandra Schwartz, and Luke Steverango. A study of one-parameter regularization methods for mathematical programs with vanishing constraints. Optimization Methods and Software, 37(2):503–545, 2022. doi:10.1080/10556788.2020.1797025.

[9] Qingjie Hu, Jiguang Wang, and Yu Chen. New dualities for mathematical programs with vanishing constraints. Annals of Operations Research, 287:233–255, 2020. doi:10.1007/s10479-019-03409-6.

[10] Alexey Feridovich Izmailov and Artur Levonovich Pogosyan. Optimality conditions and newton-type methods for mathematical programs with vanishing constraints. Computational Mathematics and Mathematical Physics, 49:1128–1140, 2009. doi:10.1134/S0965542509070069.

[11] RA Jabr. Solution to economic dispatching with disjoint feasible regions via semidefinite programming. IEEE Transactions on power systems, 27(1):572–573, 2011. doi:10.1109/TPWRS.2011.2166009.

[12] Michael N Jung, Christian Kirches, and Sebastian Sager. On perspective functions and vanishing constraints in mixed-integer nonlinear optimal control. In Facets of Combinatorial Optimization*: Festschrift for Martin Grötschel*, pages 387–417. Heidelberg: Springer Berlin Heidelberg, Berlin, 2013. doi:10.1007/978-3-642-38189-8_16.

[13] Christian Kirches, Andreas Potschka, Hans Georg Bock, and Sebastian Sager. A parametric active set method for quadratic programs with vanishing constraints. Technical Report, 2012.

[14] Vivek Laha, Rahul Kumar, Harsh Narayan Singh, and SK Mishra. On minimax programming with vanishing constraints. In Optimization, Variational Analysis and Applications*: IFSOVAA-2020, Varanasi, India, February 2–4*, pages 247–263, Singapore, 2021. Springer Singapore. doi:10.1007/978-981-16-1819-2_11.

[15] Vivek Laha and Lalita Pandey. On mathematical programs with equilibrium constraints under data uncertainty. In International Conference on Nonlinear Applied Analysis and Optimization, pages 283–300. Springer, 2021.

[16] Vivek Laha, Vinay Singh, Yogendra Pandey, and SK Mishra. Nonsmooth mathematical programs with vanishing constraints in banach spaces. In High-Dimensional Optimization and Probability: With a View Towards Data Science, pages 395–417. Springer, 2022. doi:10.1007/978-3-031-00832-0_13.

[17] Gue Myung Lee and Pham Tien Son. On nonsmooth optimality theorems for robust optimization problems. Bulletin of the Korean Mathematical Society, 51(1):287–301, 2014. doi:10.4134/BKMS.2014.51.1.287.

[18] Olvi L Mangasarian and Stan Fromovitz. The fritz john necessary optimality conditions in the presence of equality and inequality constraints. Journal of Mathematical Analysis and applications, 17(1):37–47, 1967. doi:10.1016/0022-247X(67)90163-1.

[19] Shashi Kant Mishra, Vinay Singh, and Vivek Laha. On duality for mathematical programs with vanishing constraints. Annals of Operations Research, 243:249–272, 2016. doi:10.1007/s10479-015-1814-8.

[20] SK Mishra, Vinay Singh, Vivek Laha, and RN Mohapatra. On constraint qualifications for multiobjective optimization problems with vanishing constraints. Optimization methods, theory and applications, pages 95–135, 2015. doi:10.1007/978-3-662-47044-2_6.

[21] Jane J Ye. Multiplier rules under mixed assumptions of differentiability and lipschitz continuity. SIAM Journal on Control and Optimization, 39(5):1441–1460, 2000. doi:10.1137/S0363012999358476.

French-German-Spanish Conference on Optimization
Gijón, June 18-21, 2024
(pp. 71-78)

On the usage of the Henstock-Kurzweil integral in infinite horizon optimal control problems

Valeriya Lykina

valeriya.lykina@b-tu.de *Brandenburg University of Technology Cottbus-Senftenberg, Germany*

Abstract

In the present paper we motivate the incorporation of a more general integral notion, namely the Henstock-Kurzweil integral, in a formulation of infinite horizon optimal control problems and investigate its impact. This results from the necessity of distinguishing between different interpretations of the improper integral objective (e.g. Lebesgue, improper Riemann etc.) which was addressed in [18]. A first result concerning sufficient optimality conditions for the new class of optimal control problems is obtained. Relations between admissible sets and optimal solutions of the new control problem and the problems involving Lebesgue or improper Riemann integrals are discussed by means of an example. The applicability of sufficient optimality conditions is also shown.

1. Introduction

The class of infinite horizon optimal control problems deserved much interest in recent time, since it has a great amount on applications in various fields, such as economics, biology, continuum mechanics. For some innovatory applications of this class of problems in drug and terror models we refer the reader to [11].

In an optimal control problem in Lagrange form one usually minimizes an integral functional of the form $\int_a^b r(t,x(t),u(t))dt$. If the integral is taken over an unbounded interval it makes an essential difference which integral notion is used. In [18] it was shown that in dependence on the used integral, i.e. Lebesgue or improper Riemann integral, the admissible set of the corresponding optimal control problem changes also. In general so does the optimal solution, cf. the same paper. The main reason for this discrepancy is that the Lebesgue integral is an "absolute" integral, i.e. a function r is Lebesgue integrable over $[0,\infty)$ if and only if its absolute value $|r|$ is Lebesgue integrable over $[0,\infty)$ as well, while the improper Riemann integral is a non-absolute one. To convince yourself in this fact consider for instance the integrand r defined by the rule $r(t)=\frac{\sin(t)}{t}$. Obviously, r itself yields a finite value of the improper Riemann integral $R\text{-}\int_0^\infty \frac{\sin(t)}{t}dt$, namely $\pi/2$, in constrast to the function $|r|$ which produces a divergent improper Riemann integral $R\text{-}\int_0^\infty \left|\frac{\sin(t)}{t}\right|dt$. However, most part of theoretical results on infinite horizon optimal control problems deal with problems involving the Lebesgue integral, while from the numerical point of view it is comfortable to use the Riemann interpretation of the integral because of the simple construction of integral sums. To the knowledge of author, by now there are no results about sufficient and necessary optimality conditions for control problems with the improper Riemann integral involved, which do not rely on a kind of absolute integrability assumption for the objective. One of the reasons for it is that the Riemann integrable functions do not build quite satisfactory functional spaces to work with. The sketched theoretical-numerical discrepancy lead to the subject of this paper.

The idea of this paper is to use a more general integration theory, namely the Henstock-Kurzweil integration, in order to formulate an optimal control problem with infinite horizon and to obtain sufficient optimality conditions via duality theory approach. For the cases, when the optimal value of the objective is not changed through the generalization of the integral notion and expending of the admissible set, the derived results on duality theory could fill the gap arised through the absence of duality theory for control problems involving the improper Riemann integral. The Henstock-Kurzweil integral, sometimes also called generalized Riemann integral or a gauge integral, is a generalization of both Lebesgue and Riemann integration theories and preserves some useful properties of Lebesgue integral being simultaneously a "non-absolute" integral, so that all Riemann integrable functions are also Henstock-Kurzweil integrable. Originally, this integral notion was introduced by a chech mathematician Jaroslav Kurzweil in last sixties of the last century while investigating differential equations with highly oscillating terms. Afterwards, an english mathematician Ralph Henstock independently rediscovered this integral and made significant contributions to its theory. For further reading on Henstock-Kurzweil integral and its properties we refer among others to [19], [15].

 ISBN 978-84-10135-30-7

The paper has the following structure. Section 2 introduces the necessary facts concerning Henstock-Kurzweil integration. In section 3 a class of infinite horizon optimal control problems arising through using the Henstock-Kurzweil integration theory is described. Section 4 is devoted to the construction of a dual problem and deriving sufficient conditions for strong duality between problems (P_{HK}) and (D_{HK}), which in turn is sufficient for optimality of the involved process. A comparison to the construction of a dual problem in case of a control problem with the Lebesgue interpretation of the integral is also given. In section 5 we show with the help of an example how the admissible set and the optimal solution of a problem of the new class can relate to those of the corresponding control problems involving the widely used Lebesgue and improper Riemann integrals. In this section we also apply the result on sufficient optimality conditions and compare dual problems obtained for all three types of discussed integrals. Section 6 contains open questions and conclusions.

2. Preliminaries

2.1. Definitions

Let us introduce $\mathbb{B}$ as a measurable set in s-dimensional Euclidean space. We denote by $\mathcal{M}^n(\mathbb{B})$, $L_p^n(\mathbb{B})$ and $C^{0,n}(\mathbb{R}^+)$ the spaces of all vector functions $x : \mathbb{B} \to \mathbb{R}^n$ with Lebesgue measurable, in the pth power Lebesgue integrable or continuous components, respectively ([9], p. 146 and pp. 285 ff.; [10], pp. 228 ff.). For $n = 1$, we suppress the superscript in the labels of the spaces. We write $[0, \infty) = \mathbb{R}^+$. In the sequel the notations $L\text{-}\int$, $R\text{-}\int$ and $HK\text{-}\int$ stand respectively for the Lebesgue, the Riemann and the Henstock-Kurzweil interpretations of the integral.

Definition 2.1 **(a)** A continuous function $\nu : \mathbb{R}^+ \to \mathbb{R}^+$ is called a *weight function*.

(b) If it additionally satisfies $L\text{-}\int_0^\infty \nu(t)\,dt < \infty$, we call it *density function.*

Definition 2.2 **(a)** By means of a weight function ν, we define for any $1 \le p < \infty$ the *weighted Lebesgue space*

$$L_p^n(\mathbb{B}, \nu) = \left\{ x \in \mathcal{M}^n(\mathbb{B}) \mid (L\text{-}\int_{\mathbb{B}} |x(t)|^p \nu(t)\,dt)^{1/p} < \infty \right\} \tag{2.1}$$

as well as

$$L_\infty^n(\mathbb{B}, \nu) = \left\{ x \in \mathcal{M}^n(\mathbb{B}) \mid ess\sup_{t \in \mathbb{B}} |x(t)\nu(t)| < \infty \right\} \tag{2.2}$$

and

(b) the *weighted Sobolev space*

$$W_p^{1,n}(\mathbb{R}^+, \nu) = \left\{ x \in \mathcal{M}^n(\mathbb{R}^+) \;\middle|\; x \in L_p^n(\mathbb{R}^+, \nu),\ \dot{x} \in L_p^n(\mathbb{R}^+, \nu) \right\} \tag{2.3}$$

(see [14], p. 11 f.), here $\dot{x}$ denotes the generalized derivative. Equipped with the norm

$$\|x\|_{W_p^{1,n}(\mathbb{R}^+,\nu)} = \|x\|_{L_p^n(\mathbb{R}^+,\nu)} + \|\dot{x}\|_{L_p^n(\mathbb{R}^+,\nu)}, \tag{2.4}$$

$W_p^{1,n}(\mathbb{R}^+, \nu)$ becomes a Banach space (this can be confirmed analogously to [14], p. 19, Theorem 3.6.).

(c) the *space of functions of bounded variation*

$$BV(\mathbb{B}) = \left\{ x \in L_1^n(\mathbb{B}) \mid Var(x, \mathbb{B}) = \sup \left\{ \int_{\mathbb{B}} x(t) div\phi(t) dt : \phi \in C_c^1(\mathbb{B}, \mathbf{R}^n),\ \|\phi\|_{L_\infty(\mathbb{B})} \le 1 \right\} < \infty \right\}$$

The following definition is leaned on [15], pp. 139 – 140 and [5], p. 626 ff. $BV(\mathbf{R}^+)$ is a non separable Banach space.

Definition 2.3 Given an interval $I = [a, b] \subseteq \mathbf{R}^+$. Then

(a) a *tagged partition* of I is a finite set of ordered pairs $\mathcal{P} = \{(t_i, I_i),\ i = 1, \dots, r\}$ such that I_i are closed subintervals having disjoint interiors, $t_i \in I_i$ and $\cup_{i=1}^r I_i = I$. The point t_i is called the *tag* associated with the interval I_i.

(b) a function $\delta : I \to \mathbf{R}^+$ is called a *gauge*.

(c) a tagged partition P is said to be *δ-fine*, if for all $i \in \{1, \dots, r\}$ the following inclusion holds: $I_i \subset (t_i - \delta(t_i), t_i + \delta(t_i))$.

Definition 2.4 **(a)** Given a tagged partition $\mathcal{P} = \{(t_i, I_i),\ i = 1, \dots, r\}$ of I, we call

$$S(f, \mathcal{P}) = \sum_{i=1}^{r} f(t_i)\mu(I_i)$$

the *Riemann sum* with respect to $\mathcal{P}$, where $\mu(\cdot)$ denotes the Lebesgue measure.

(b) A function $f : I \to \mathbf{R}$ is called *Henstock-Kurzweil integrable* or shortly *HK-integrable*, if there is an $A \in \mathbf{R}$ so that for all $\epsilon > 0$ there exists a gauge δ on I so that for every δ-fine partition P of I $|S(f, P) - \mathrm{A}| < \epsilon$ holds. We call the number A the *Henstock-Kurzweil integral* of f over I and write $A = HK\text{-}\int_I f = HK\text{-}\int_a^b f$.

(c) We denote by $HK(I)$ the space of all HK-integrable functions on I. This space can be equipped with the Alexiewicz semi-norm, $\|f\|_A = \sup_{t \in I} \left| HK\text{-}\int_a^t f \right|$, which induces the normed space of equivalence classes.

Definition 2.5 As a kind of analogon of the weighted Lebesgue- and Sobolev spaces defined in Definition 2.2 we introduce

(a) by means of a weight function v for a measurable set $\mathbb{B}$ the weighted space $HK^n(\mathbb{B}, v) = \left\{ x \in \mathcal{M}^n(\mathbb{B}) \mid HK\text{-}\int_{\mathbb{B}} x(t)\, v(t)\, dt < \infty \right\}$; whereas if $v(t) \equiv 1$, we denote the space by $HK^n(\mathbb{B})$.

(b) the weighted space $HK^{1,n}(\mathbb{B}, v) = \left\{ x \in HK^n(\mathbb{B}, v) \;\middle|\; \dot{x} \in HK^n(\mathbb{B}, v) \right\}$.

Lemma 2.6 *Let a density function v be given. Any linear, continuous functional $\varphi : HK(\mathbb{R}^+, v) \to \mathbb{R}$ can be represented by a function $y \in BV(\mathbb{R}^+)$:*

$$\langle \varphi, x \rangle = HK\text{-}\int_0^\infty y(t)\, x(t)\, v(t)\, dt \quad \forall\, x \in HK(\mathbb{R}^+, v)\,. \tag{2.5}$$

We can apply [10], p. 287, Theorem 3.2, since the measure generated by the density function v is σ-finite on $\mathbb{R}^+$.

The next analogon of Hölder's inequality is taken from [24], p. 62.

Theorem 2.7 *(Hölder's inequality for distributions)*
Let $f \in \mathcal{A}_C = \left\{ f \in \mathcal{D}' \mid \exists\, F \in C^0(\bar{\mathbf{R}})$ with $F(-\infty) = 0,\ F' = f \right\}$. For $g \in BV(\mathbf{R})$ it holds then the inequality

$$\left| HK\text{-}\int_{-\infty}^{\infty} fg \right| \le 2\|f\|_A \cdot \|g\|_{BV}. \tag{2.6}$$

2.2. Infinite horizon optimal control problem

It is considered the infinite horizon control problem of minimizing the integral objective

$$J_\infty^{HK}(x, u) = HK\text{-}\int_0^\infty r(\cdot, x(\cdot), u(\cdot))\tilde{v}(\cdot) \to \min! \tag{2.7}$$

with respect to all pairs satisfying the following constraints:

$$(x, u) \in HK^{1,n}(\mathbb{R}^+, v) \times HK^m(\mathbb{R}^+, v)\,; \tag{2.8}$$

$$\dot{x}(t) = f(t, x(t), u(t)) \quad \text{a.e. on } \mathbb{R}^+\,; \tag{2.9}$$

$$x(0) = x_0\,; \tag{2.10}$$

$$u(t) \in U \quad \text{a.e. on } \mathbb{R}^+ \tag{2.11}$$

Hereby U denotes a compact convex subset of $\mathbb{R}^m$, v and $\tilde{v}$ are weight functions due to the Definition 2.1. The functions x and u are called the state and the control function respectively. The integral in (2.7) is understood in Henstock-Kurzweil sense. We refer to problem (2.7) – (2.11) as to the problem (P_{HK}). We now introduce

Assumption 1 *The function $r(t, \xi, v)$ is continuous in t, continuously differentiable in ξ and v and convex in v.*

Definition 2.8 **(a)** A pair (x, u) is called *admissible* for the problem (P_{HK}), if it satisfies the conditions (2.8) – (2.11) and the Henstock-Kurzweil integral in (2.7) exists and has a finite value.

(b) An admissible pair (x^*, u^*) is called a *global optimal solution* of the problem (P_{HK}), if for any admissible pair (x, u) the inequality $J_\infty^{HK}(x^*, u^*) \le J_\infty^{HK}(x, u)$ holds.

We also introduce the following problems with the Lebesgue and improper Riemann interpretation of the integral in the objective:

$$(P_L): \begin{cases} J_\infty^L(x,u) = L\text{-}\int_0^\infty r(t,x(t),u(t))\tilde{v}(t)dt \to \min! \\ (x,u) \in W_2^{1,1}(\mathbb{R}^+, v) \times L_2^1(\mathbb{R}^+, v) \\ \dot{x}(t) = f(t,x(t),u(t)) \quad a.e.\ on\ \ \mathbb{R}^+ \\ u(t) \in U \quad a.e.\ on\ \ \mathbb{R}^+ \\ x(0) = x_0 \end{cases}; (P_R): \begin{cases} J_\infty^R(x,u) = R\text{-}\int_0^\infty r(t,x(t),u(t))\tilde{v}(t)dt \to \min! \\ (x,u) \in W_2^{1,1}(\mathbb{R}^+, v) \times L_2^1(\mathbb{R}^+, v) \\ \dot{x}(t) = f(t,x(t),u(t)) \quad a.e.\ on\ \ \mathbb{R}^+ \\ u(t) \in U \quad a.e.\ on\ \ \mathbb{R}^+ \\ x(0) = x_0 \end{cases}.$$

3. Dual problem and sufficient optimality conditions

Generally speaking we call a problem **(D)**: $\max_{y \in Y} g(y)$ *weakly dual* to the problem **(P)**: $\min_{x \in X} f(x)$, if for all $x \in X$ and for all $y \in Y$ the inequality $f(x) \ge g(y)$ holds. If there are some $x^* \in X$ and $y^* \in Y$ such that the equation $f(x^*) = g(y^*)$ is satisfied, we call these two problems *strong dual* to each other. Similar idea of duality can be applied to optimal control problems and specifically to the problems with infinite horizon, see [13], [21], [22], [16]. The special construction scheme for dual problems used in cited papers is due to R. Klötzler, [13]. We follow this scheme in the present paper as well. Before formulating the main result we need some auxiliary result.

Lemma 3.1 *Let* $\mathbb{R}^+ = \bigcup_{k=0}^{r-1} [\tau_k, \tau_{k+1})$, $\tau_0 := 0$, $\tau_r := \infty$ *be a partition of the half-axis* $\mathbb{R}^+$ *into disjoint intervals. Furthermore, let* (x^*, u^*) *be an admissible pair of* (P_{HK}) *and* $S : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}$ *be a function of the form*

$$S(t,\xi) = a(t) + y(t)^T(\xi - x^*(t)), \tag{3.1}$$

having $a \in HK^{1,1}(\mathbb{R}^+)$; $\frac{y}{v}, \frac{\dot{y}}{v} \in BV(\mathbb{R}^+)$, $y \in C^1(\tau_k, \tau_{k+1})$ *for all* $k = 1, \dots, r-1$. *Then for any* $x \in HK^{1,n}(\mathbb{R}^+, v)$ *with* $x(0) = x_0$ *we have:*

$$\lim_{T\to\infty} S(T, x(T)) = 0, \tag{3.2}$$

$$HK\text{-}\int_0^\infty \frac{d}{dt} S(\cdot, x(\cdot)) = -S(0, x_0) + \sum_{k=1}^{r-1} (S(\tau_k - 0, x(\tau_k)) - S(\tau_k + 0, x(\tau_k))). \tag{3.3}$$

Proof: *We estimate* $\left|HK\text{-}\int_0^\infty S(\cdot, x(\cdot))\right| \le \left|HK\text{-}\int_0^\infty a(\cdot)\right| + \left|HK\text{-}\int_0^\infty y(\cdot)^T(x(\cdot) - x^*(\cdot))\right|$ *and applying the Hölder's inequality to the last term obtain*

$$\left|HK\text{-}\int_0^\infty S(\cdot, x(\cdot))\right| \le \|a\|_{HK^1(\mathbb{R}^+)} + 2\left\|\frac{y}{v}\right\|_{BV(\mathbb{R}^+)} \cdot \|x - x^*\|_{HK^1(\mathbb{R}^+, v)} < \infty. \tag{3.4}$$

From the finiteness of the integral $HK\text{-}\int_0^\infty S(\cdot, x(\cdot))$ *we conclude (3.2), since*

$$\lim_{T\to\infty} HK\text{-}\int_0^T S(\cdot, x(\cdot)) = \lim_{T\to\infty} \left(HK\text{-}\int_0^{T-1} S(\cdot, x(\cdot)) + HK\text{-}\int_{T-1}^T S(\cdot, x(\cdot)) \right) = \lim_{T\to\infty} HK\text{-}\int_0^T S(\cdot, x(\cdot)) + \lim_{\tau\to\infty} S(\tau, x(\tau)),$$

whereby $\tau \in [T-1, T]$. *Similarly, one shows the existence and finiteness of the integral* $HK\text{-}\int_0^\infty \frac{dS(\cdot,x(\cdot))}{dt}$ *using the expression* $\frac{dS(t,x(t))}{dt} = \dot{a}(t) + \dot{y}(t)^T(x(t) - x^*(t)) + y(t)^T(\dot{x}(t) - \dot{x}^*(t))$ *for the linear function S given in (3.1):*

$$\left|HK\text{-}\int_0^\infty \frac{dS(\cdot, x(\cdot))}{dt}\right| \le \|\dot{a}\|_{HK^1(\mathbb{R}^+)} + 2\left\|\frac{\dot{y}}{v}\right\|_{BV(\mathbb{R}^+)} \cdot \|x - x^*\|_{HK^1(\mathbb{R}^+, v)} + 2\left\|\frac{y}{v}\right\|_{BV(\mathbb{R}^+)} \cdot \|\dot{x} - \dot{x}^*\|_{HK^1(\mathbb{R}^+, v)} < \infty.$$

The condition (3.3) can now be derived by means of (3.2). ∎

We introduce the Hamiltonian as $\mathcal{H}(t,\xi,\eta) = \sup_{v \in U} H(t,\xi,v,\eta)$ with $H(t,\xi,v,\eta) = -r(t,\xi,v) + \frac{1}{\tilde{v}(t)} < \eta, f(t,\xi,v) >$ where H represents the Pontryagin's function. Furthermore, we define the reachable set

$$R(t) = \left\{ \xi \in \mathbb{R}^n \;\middle|\; \begin{array}{c} \exists\,(x,u) \in HK^{1,n}(\mathbb{R}^+, v) \times HK^m(\mathbb{R}^+, v) \text{ mit} \\ \xi = x(t),\ \dot{x}(t) = f(t,x(t),u(t)),\ u(t) \in U,\ x(0) = x_0 \end{array} \right\} \tag{3.5}$$

and

$$Y_{HK} = \left\{ S : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R} \;\middle|\; \begin{array}{c} S(t,\xi) \;=\; a(t) + y(t)^T(\xi - x^*(t)) \\ \\ a \in HK^{1,1}(\mathbb{R}^+),\ \frac{y}{\nu}, \frac{\dot{y}}{\nu} \in BV(\mathbb{R}^+), \\ y \in C^0[\tau_k, \tau_{k+1}] \cap C^1(\tau_k, \tau_{k+1}),\ k = 1, \dots, r-1 \\ \\ \frac{1}{\tilde{\nu}(t)}\partial_t S(t,\xi) + \mathcal{H}(t,\xi,\partial_\xi S(t,\xi)) \le 0 \\ \forall (t,\xi) \in \left(\mathbb{R}^+ \backslash \{\tau_1, \dots, \tau_{r-1}\}\right) \times R(t) \end{array} \right\}. \tag{3.6}$$

Using the scheme for constructing dual problems described in [13] we construct a problem (D_{HK}) and prove

Theorem 3.2 *(Weak duality relation)*
Let a problem (P_{HK}) be given. Then for the problem (D_{HK}):

$$g_\infty(S) \;:=\; -S(0,x_0) + \inf_{\beta \in \Gamma}\left\{\sum_{k=1}^{r-1}(S(\tau_k - 0, \beta_k) - S(\tau_k + 0, \beta_k))\right\} \to \sup!\ \text{ w.r.t. } S \in Y_{HK} \tag{3.7}$$

$$\Gamma \;=\; \left\{\beta = (\beta_1, \dots, \beta_{r-1}) \in \mathbb{R}^{(r-1)n} \,|\, \beta_k \in R(\tau_k)\right\} \tag{3.8}$$

the following weak duality relation is true:

$$\inf(P_{HK}) \ge \sup(D_{HK}). \tag{3.9}$$

Definition 3.3 A linear in ξ ansatz S is called admissible for the problem (D_{HK}), if the inclusion $S \in Y_{HK}$ holds.

Proof: *Let (x,u) be admissible for (P_{HK}) and S be admissible for (D_{HK}), i.e. $S \in Y_{HK}$. Then we have the following estimate*

$$\begin{aligned} J_\infty^{HK}(x,u) &= HK\text{-}\int_0^\infty r(\cdot,x(\cdot),u(\cdot))\tilde{\nu}(\cdot) = \\ &= HK\text{-}\int_0^\infty \left(-H(\cdot,x(\cdot),u(\cdot),\partial_\xi S(\cdot,x(\cdot)))\right)\tilde{\nu}(\cdot) + HK\text{-}\int_0^\infty \left(\frac{\partial_\xi S(\cdot,x(\cdot))}{\tilde{\nu}(\cdot)} g(\cdot,x(\cdot),u(\cdot))\right)\tilde{\nu}(\cdot) \\ &= HK\text{-}\int_0^\infty \left(-H(\cdot,x(\cdot),u(\cdot),\partial_\xi S(\cdot,x(\cdot))) - \frac{\partial_t S(\cdot,x(\cdot))}{\tilde{\nu}(\cdot)}\right)\tilde{\nu}(\cdot) + HK\text{-}\int_0^\infty \left(\frac{\partial_t S(\cdot,x(\cdot))}{\tilde{\nu}(\cdot)} + \frac{\partial_\xi S(\cdot,x(\cdot))}{\tilde{\nu}(\cdot)}\dot{x}(\cdot)\right)\tilde{\nu}(\cdot) \\ &\ge -HK\text{-}\int_0^\infty \left(\mathcal{H}(\cdot,x(\cdot),\partial_\xi S(\cdot,x(\cdot))) + \frac{\partial_t S(\cdot,x(\cdot))}{\tilde{\nu}(\cdot)}\right)\tilde{\nu}(\cdot) + HK\text{-}\int_0^\infty \left(\partial_t S(\cdot,x(\cdot)) + \partial_\xi S(\cdot,x(\cdot))\dot{x}(\cdot)\right). \end{aligned}$$

Taking the Hamilton-Jacobi inequality, which has to be fulfilled by S, see the definition of the set Y_{HK}, and Lemma 3.1 into account one arrives at

$$\begin{aligned} &J_\infty^{HK}(x,u) \ge -HK\text{-}\int_0^\infty \sup_{\xi \in R(t)}\left\{\left(\mathcal{H}(\cdot,\xi,\partial_\xi S(t,\xi)) + \frac{\partial_t S(\cdot,\xi)}{\tilde{\nu}(\cdot)}\right)\right\}\tilde{\nu}(\cdot) \\ &+HK\text{-}\int_0^\infty \left(\partial_t S(\cdot,x(\cdot)) + \partial_\xi S(\cdot,x(\cdot))\dot{x}(\cdot)\right) \ge HK\text{-}\int_0^\infty \frac{d}{dt}S(\cdot,x(\cdot)) = \lim_{T\to\infty} HK\text{-}\int_0^T \frac{d}{dt}S(\cdot,x(\cdot)) \\ &= \lim_{T\to\infty} S(T,x(T)) - S(0,x(0)) + \sum_{k=1}^{r-1}(S(\tau_k - 0, x(\tau_k)) - S(\tau_k + 0, x(\tau_k))) \\ &= -S(0,x_0) + \sum_{k=1}^{r-1}(S(\tau_k - 0, x(\tau_k)) - S(\tau_k + 0, x(\tau_k))) \ge -S(0,x_0) + \inf_{\beta\in\Gamma}\left\{\sum_{k=1}^{r-1}(S(\tau_k - 0, \beta_k) - S(\tau_k + 0, \beta_k))\right\} \end{aligned}$$

which closes the proof. ∎

Remark 3.4 As we can see, the proper decision variable in the dual problem (D_{HK}) is (a,y), but we use $S \in Y_{HK}$ for simplicity. Furthermore, the component y of the dual process S belongs to a Banach space, whereas the state trajectory itself does not. This can be advantageous in deriving existence results for the primal problem (P_{HK}).

Theorem 3.5 *An admissible pair (x^*, u^*) is a global minimizer of (P_{HK}), if there exists an admissible S^* for (D_{HK}), such that the following conditions are fulfilled for almost all $t > 0$:*

(M) $$\mathcal{H}(t, x^*(t), \partial_\xi S^*(t, x^*(t))) = H(t, x^*(t), u^*(t), \partial_\xi S^*(t, x^*(t)))$$

(HJ) $$\frac{1}{\nu(t)} S_t^*(t, x^*(t)) + \mathcal{H}(t, x^*(t), \partial_\xi S(t, x^*(t))) = 0$$

(B) $$\inf_{\zeta \in \mathbb{R}^{n(r-1)}} \left\{ \sum_{k=1}^{r-1} (S^*(\tau_k - 0, \zeta_k) - S^*(\tau_k + 0, \zeta_k)) \right\} = \sum_{k=1}^{r-1} (S^*(\tau_k - 0, x^*(\tau_k)) - S^*(\tau_k + 0, x^*(\tau_k)))$$

Proof: *follows immediately from Theorem 3.2 and the assumptions of this theorem.* ∎

To compare the constructed dual problem (D_{HK}) with (D_L) for the corresponding control problem involving the Lebesgue integral, cf. [16], p. 107 ff., we notice that the dual ansatz S has to belong to the set

$$Y_L = \left\{ S : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R} \;\middle|\; \begin{array}{c} S(t,\xi) = a(t) + y(t)^T(\xi - x^*(t)) \\ \\ a \in W_1^1([\tau_{r-1}, \infty)) \cap C^0[\tau_k, \tau_{k+1}] \cap C^1(\tau_k, \tau_{k+1}) \\ y \in W_q^{1,n}([\tau_{r-1}, \infty), \nu^{1-q}) \cap C^0[\tau_k, \tau_{k+1}] \cap C^1(\tau_k, \tau_{k+1}) \\ i = 0, \dots, r-1 \\ \\ \frac{1}{\nu(t)} \partial_t S(t,\xi) + \mathcal{H}(t, \xi, \partial_\xi S(t,\xi)) \le 0 \\ \forall (t,\xi) \in \left(\mathbb{R}^+ \backslash \{\tau_1, \dots, \tau_{r-1}\}\right) \times R(t) \end{array} \right\}, \tag{3.10}$$

where one has $(a, y) \in W_1^1([\tau_{r-1}, \infty)) \times W_q^{1,n}([\tau_{r-1}, \infty), \nu^{1-q})$ instead of $a \in HK^{1,1}(\mathbb{R}^+)$ and $\frac{y}{\nu}, \frac{\dot{y}}{\nu} \in BV(\mathbb{R}^+)$ in case of the problem (D_{HK}). Apart from that the dual problems (D_{HK}) and (D_L) have the same form.

4. Example and application of the theoretical result

Example 4.1 The following example was considered in [23], p. 464, although with Lebesgue integral involved in the objective function and weighted Lebesgue- and Sobolev spaces as the spaces for the state and control functions respectively. We now change the interpretation of the integral in the objective to the Henstock-Kurzweil integral up and consider the weighted Henstock-Kurzweil spaces as the state- and control spaces respectively. Thus, with $\nu(t) = e^{-t}$ and $r(t) = \begin{cases} \frac{\sin t}{t} & , \quad 2k\pi \le t \le (2k+1)\pi \\ 2\frac{\sin t}{t} & , \quad (2k+1)\pi \le t \le (2k+2)\pi \end{cases}$, we obtain the problem

$$(P_{HK}^1): \begin{cases} J_\infty^{HK}(x,u) = HK\text{-}\int_0^\infty r(\cdot)u(\cdot) \to \max! \\ (x,u) \in HK^{1,1}(\mathbb{R}^+, \nu) \times HK^1(\mathbb{R}^+, \nu) \\ \dot{x}(t) = u(t) \quad a.e.\ on\ \mathbb{R}^+ \\ u(t) \in [\frac{1}{2}; 1] \quad a.e.\ on\ \mathbb{R}^+ \\ x(0) = 0 \end{cases} . \tag{4.1}$$

For the possibility of comparison we formulate both the corresponding problems (P_L^1) and (P_R^1) and denote their admissible sets by $\mathcal{A}_L^1$ and $\mathcal{A}_R^1$ respectively:

$$(P_L^1): \begin{cases} J_\infty^L(x,u) = L\text{-}\int_0^\infty r(t)u(t)dt \to \max! \\ (x,u) \in W_2^{1,1}(\mathbb{R}^+, \nu) \times L_2^1(\mathbb{R}^+, \nu) \\ \dot{x}(t) = u(t) \quad a.e.\ on\ \mathbb{R}^+ \\ u(t) \in [\frac{1}{2}; 1] \quad a.e.\ on\ \mathbb{R}^+ \\ x(0) = 0 \end{cases} ; \quad (P_R^1): \begin{cases} J_\infty^R(x,u) = R\text{-}\int_0^\infty r(t)u(t)dt \to \max! \\ (x,u) \in W_2^{1,1}(\mathbb{R}^+, \nu) \times L_2^1(\mathbb{R}^+, \nu) \\ \dot{x}(t) = u(t) \quad a.e.\ on\ \mathbb{R}^+ \\ u(t) \in [\frac{1}{2}; 1] \quad a.e.\ on\ \mathbb{R}^+ \\ x(0) = 0 \end{cases}$$

Assertion 1 **(i)** *The admissible set $\mathcal{A}_{HK}^1$ of the problem (P_{HK}^1) satisfies the relation $\mathcal{A}_{HK}^1 \supset (\mathcal{A}_L^1 \cup \mathcal{A}_R^1)$.*

(ii) *For the optimal values it holds* $\sup(P_{HK}^1) = \max\{\sup(P_L^1), \sup(P_R^1)\} = \sup(P_R^1)$

Proof: *(i): The inclusion $\mathcal{A}_{HK}^1 \supseteq (\mathcal{A}_L^1 \cup \mathcal{A}_R^1)$ is clear, since if a function is integrable in either Lebesgue or Riemann sense, it is also integrable in Henstock-Kurzweil sense. More interesting is to prove the strict inclusion,*

(P_L^1)	(P_R^1)	(P_{HK}^1)
(D_L^1) has empty admissible set	(D_R^1) cannot be formulated	(D_{HK}^1) is strong dual to (P_{HK}^1) with $a(t) = \frac{\pi}{2} - HK\text{-}\int_0^t \frac{\sin(\cdot)}{(\cdot)}$, $y(t) \equiv 0$

Tab. 1 Comparison of dual problems

means the existence of a pair $(x,u) \in \mathcal{A}_{HK}^1 \backslash (\mathcal{A}_L^1 \cup \mathcal{A}_R^1)$. *We recall that* $\mathcal{A}_L^1 = \emptyset$ *and the optimal control for the Riemann case is*

$$u_R^*(t) = \begin{cases} 1 & , \quad t \in [2k\pi, (2k+1)\pi) \\ 1/2 & , \quad t \in [(2k+1)\pi, (2k+2)\pi) \end{cases} . \tag{4.2}$$

We construct a function $u_{HK}(\cdot)$, *for which the set of discontinuity points has a positive measure. To this end we change up the function* $u_R^*(\cdot)$ *on the Cantor set C of measure* $\mu(C) > 0$ *(For algorithm of construction of Cantor sets of positive measure we refer to [2], p. 236):*

$$u_{HK}(t) = \begin{cases} \begin{cases} 1 & , \quad t \in \bar{C} \cap [2k\pi, (2k+1)\pi) \\ 1/2 & , \quad t \in C \cap [2k\pi, (2k+1)\pi) \end{cases} \\ \begin{cases} 1/2 & , \quad t \in \bar{C} \cap [(2k+1)\pi, (2k+2)\pi) \\ 1 & , \quad t \in C \cap [(2k+1)\pi, (2k+2)\pi) \end{cases} \end{cases} . \tag{4.3}$$

The so constructed suboptimal control function $u_{HK}(\cdot)$ *is integrable only in Henstock-Kurzweil sense, but not in the Riemann sense, sinse the discontinuity points form a set of positive measure. With the corresponding state function* $x_{HK}(t) := x_0 + HK\text{-}\int_0^t u_{HK}$ *we arrive at the pair* $(x_{HK}, u_{HK}) \in \mathcal{A}_{HK}^1 \backslash (\mathcal{A}_L^1 \cup \mathcal{A}_R^1)$ *which proves one part of the assertion.*

(ii): For a pair $(x,u) \in (\mathcal{A}_L^1 \cup \mathcal{A}_R^1)$ *it holds* $J_\infty^{HK}(x,u) = \max\{J_\infty^L, J_\infty^R\}$ *due to the inclusion proved in* (i) *and to the following implications. If the Lebesgue integral* $J_\infty^L(x,u)$ *exists, then the Henstock-Kurzweil integral* $J_\infty^{HK}(x,u)$ *exists also, cf. Section 2, and the equality* $J_\infty^{HK}(x,u) = J_\infty^L(x,u)$ *is valid. If the improper Riemann integral* $J_\infty^R(x,u)$ *exists, then* $J_\infty^{HK}(x,u) = J_\infty^R(x,u)$ *is true as well.*

Now let $(x,u) \in \mathcal{A}_{HK}^1 \backslash (\mathcal{A}_L^1 \cup \mathcal{A}_R^1)$. *It means that* (x,u) *is admissible neither for* (P_L^1) *nor for* (P_R^1). *Moreover, we know that* $\mathcal{A}_L^1 = \emptyset$. *Furthermore, assume that* $J_\infty^{HK}(x,u) > J_\infty^R(x_R^*, u_R^*)$, *whereby* $(x_R^*, u_R^*)_\infty \neq (x,u)$ *denotes the optimal solution of* (P_R^1). *From* $(x,u) \notin \mathcal{A}_R^1$ *we conclude that either the integral* $R\text{-}\int_0^\infty r(t)u(t)dt$ *does not exist, which means the function* $r(\cdot)u(\cdot)$ *has the set of discontinuity points of positive measure, or the control function* $u(\cdot)$ *itself is not Riemann integrable so that the state equation is violated. Taking the sign of* $\frac{\sin t}{t}$ *and the control constraint* $u(t) \in [\frac{1}{2}, 1]$ *into account we estimate for a* $k \in \mathbb{N}$

$$HK\text{-}\int_{2k\pi}^{(2k+1)\pi} \frac{\sin(\cdot)}{(\cdot)} u(\cdot) \leq HK\text{-}\int_{2k\pi}^{(2k+1)\pi} \frac{\sin(\cdot)}{(\cdot)} \quad \text{and} \quad HK\text{-}\int_{(2k+1)\pi}^{(2k+2)\pi} 2\frac{\sin(\cdot)}{(\cdot)} u(\cdot) \leq HK\text{-}\int_{(2k+1)\pi}^{(2k+2)\pi} 2\frac{\sin(\cdot)}{(\cdot)} \frac{1}{2}$$

Summarizing the situation on $\mathbf{R}^+$ *we can say that the functional* J_∞^{HK} *attains the maximal value for* $u(t) = u_R^*(t)$ *for all* $t \in \mathbf{R}^+$. *And each deviation of* $u(\cdot)$ *from* $u_R^*(\cdot)$ *on a set of positive Lebesgue measure leads to a suboptimal value of the functional. Consequently,* $(x,u) = (x_R^*, u_R^*)$ *in the sense of equivalence classes, i.e.* $(x(t), u(t)) = (x_R^*(t), u_R^*(t))$ *almost everywhere. This completes the proof.* ■

Solution to Example 1: Rewriting the problem (P_{HK}^1) as a minimization problem and applying Theorem 3.5 with $a(t) = \frac{\pi}{2} - HK\text{-}\int_0^t \frac{\sin(\cdot)}{(\cdot)}$, $y(t) \equiv 0$ and the linear dual variable $S^*(t,\xi) = a(t) + y(t)(\xi - x^*(t))$, we obtain the strong duality between the problem (P_{HK}^1) and the problem (D_{HK}^1), constructed according to Theorem 3.2. The conditions of Theorem 3.5 are satisfied along the process (x_R^*, u_R^*). Therefore, the process $(x_{HK}^*, u_{HK}^*) = (x_R^*, u_R^*)$ is a global optimal solution for the problem (P_{HK}^1). The dual problem (D_L^1) to the problem (P_L^1) has an empty admissible set, since the only possible candidate for the component $a(\cdot)$, which could lead to the strong duality between these two problems, is defined by the rule $a(t) = \frac{\pi}{2} - L\text{-}\int_0^t \frac{\sin\tau}{\tau} d\tau$ and is not admissible because of the infinite value of $a(\infty)$. For the problem (P_R^1) no dual problem can be constructed due to the used duality scheme. The summary of the comparison of dual problems (D_L^1), (D_R^1) and (D_{HK}^1) is given in Table 1.

5. Conclusions and open questions

The incorporation of the Henstock-Kurzweil integral into the setting of an optimal control problem allows to develop a satisfactory theory for problems involving a non-absolute integral in the objective, particularly the duality theory which is missed for optimal control problems with an improper Riemann integral. This opens up a new research field, since the questions of existence, necessary optimality conditions, numerical solutions etc. of this kind of problems are still uncleared.

Acknowledgements

The author was partially supported by DFG Grant LY149/2-3.

References

[1] Adams, R.A. (1978). *Sobolev Spaces*, Academic Press, San Diego.

[2] Appell, J. (2009). *Analysis in Beispielen und Gegenbeispielen*, Springer, Heidelberg.

[3] Arada, N.; Raymond. J.P. (2000). Optimal control with mixed control-state constraints. *Journal on Control and Optimization* **39**, Vol. 5, 1391 – 1407.

[4] Aseev, S. M.; Kryazhimskii, A. V. (2007). *The Pontryagin Maximum Principle and Optimal Economic Growth Problems*, Proceedings of Steklov Institute of Mathematics, Vol. **257**, Moscow.

[5] Bartle, R. (1996). Return to the Riemann integral. *Amer. Math. Monthly* **103**, No. 8, 625 – 632.

[6] Blanchard, P.; Brüning, E. (1982). *Direkte Methoden der Variationsrechnung*, Springer-Verlag, Wien, New York.

[7] Blot, J.; Hayek, N. (1996). Second order necessary conditions for the infinite-horizon variational problems. *Mathematics of Operations Research* **21**, 979 – 990.

[8] Cartigny, P.; Michel, P. (2003). On a sufficient transversality condition for infinite horizon optimal control problems. *Automatica* **39**, 1007 – 1010.

[9] Dunford N.; Schwartz, J. T. (1988). *Linear Operators. Part I: General Theory.* Wiley-Interscience; New York etc.

[10] Elstrodt, J. (1996). *Maß- und Integrationstheorie*, Springer, Berlin, 1996.

[11] Feichtinger, G.; Hartl, R. F. (1986). *Optimale Kontrolle ökonomischer Prozesse.* de Gruyter; Berlin - New York.

[12] Grass, D.; Caulkins, J. P.; Feichtinger, G.; Tragler, G.; Behrens, D. A. (2008). *Optimal Control of Nonlinear Processes With Applications in Drugs, Corruption, and Terror.* Springer; Berlin - Heidelberg.

[13] Klötzler, R. (1979). On a general conception of duality in optimal control. In: *Equadiff IV. Proceedings of the Czechoslovak Conference on Differential Equations and their Applications held in Prague, August 22 – 26, 1977.* (Fábera, J. (Ed.)), 189 – 196, Springer, Berlin (Lecture Notes in Mathematics **703**).

[14] Kufner, A. (1985). *Weighted Sobolev Spaces.* John Wiley & Sons; Chichester etc.

[15] Kurtz, D. S.; Swartz, C. W. (2005). *Theories of Integration. The Integrals of Riemann, Lebesgue, Henstock-Kurzweil, and McShane.* World Scientific; New Jersey-London-Singapore-etc.

[16] Lykina, V. (2010). *Beiträge zur Theorie der Optimalsteuerungsprobleme mit unendlichem Zeithorizont.* PhD Thesis, Cottbus.

[17] Lykina, V.; Pickenhain, S.; Wagner, M. (2008). On a resource allocation model with infinite horizon. *Applied Mathematics and Computation* **204**, pp. 595 – 601.

[18] Lykina, V.; Pickenhain, S.; Wagner, M. (2008). Different interpretations of the improper integral objective in an infinite horizon control problem. *Mathematical Analysis and Applications* **340**, pp. 498 – 510.

[19] Lee Peng-Yee (1989). *Langzhou Lectures on Henstock Integration*, World Scientific.

[20] Magill, M. J. P. (1982). Pricing infinite horizon programs. *J. Math. Anal. Appl.* **88**, 398 – 421.

[21] Pickenhain, S.; Lykina, V. (2006). Sufficiency conditions for infinite horizon optimal control problems. In: *Recent Advances in Optimization*. (Seeger, A. (Ed.)), (*Lecture Notes in Economics and Mathematical Systems* **563**), 217 – 232, Springer, Berlin etc.

[22] Pickenhain, S. (1998). Duality in optimal control with first-order differential equations. *Preprint Reihe Mathematik.* **M-08/1998**, BTU Cottbus.

[23] Pickenhain, S.; Lykina, V.; Wagner, M. (2008). On the lower semicontinuity of functionals involving Lebesgue or improper Riemann integrals in infinite horizon optimal control problems. *Control and Cybernetics* **37**, No. 2, pp. 451 – 468.

[24] Talvila, E. (2008). The distributional Denjoy integral. *Real Analysis Exchange* **33**, pp. 51 – 82.

[25] Werner, D. (1995). *Funktionalanalysis.* Springer, Berlin – Heidelberg – New York.

FRENCH-GERMAN-SPANISH CONFERENCE ON OPTIMIZATION
Gijón, June 18-21, 2024
(pp. 79-86)

Double control problem: domains and coefficients for elliptic equations

Juan Casado-Díaz[1], **Manuel Luna-Laynez**[2], **Faustino Maestre**[3]

1. jcasadod@us.es Universidad de Sevilla, Spain
2. mllynez@us.es Universidad de Sevilla, Spain
3. fmaestre@us.es Universidad de Sevilla, Spain

Abstract

In this work we are interested in a bi-optimal control problem for a linear elliptic state equation with homogeneous boundary Dirichlet condition. The two controls variables correspond to the coefficient of the diffusion term of the equation and the open set where the it is posed. From the practical point of view, this problem can be interpreted as finding materials from the mixture of other ones with different diffusion properties and on optimal shape. We analyze a relaxation process, optimality conditions, and finally we provide a numerical algorithm and we show some numerical experiments.

1. Introduction

Let Ω be a bounded open set of $\mathbb{R}^N$ considered as the domain of reference, a typical optimal design problem consists in finding the optimal layout of two materials in order to minimize a certain cost functional ([1], [11], [14]). In this sense, in the case of two isotropic materials with diffusion constants $0 < \alpha < \beta$ the problem can be formulated from the mathematical point of view:

$$\min_{\omega\subset\Omega \text{ measurable}} \int_\omega F(x,u)\,dx$$
$$\begin{cases} -\mathrm{div}\big((\alpha\chi_\omega + \beta\chi_{\Omega\setminus\omega})\nabla u\big) = f \ \text{ in } \Omega \\ u = 0 \ \text{ on } \partial\Omega \end{cases} \tag{1.1}$$

where f is a given source. The control variable $\omega \subset \Omega$ measurable determines where the material α is placed.

Another typical problem in optimal design appears when we only dispose of one material and the control variable corresponds to the place where the material is or not posed, i.e., the control variable determines the shape of the optimal domain $\omega \subset \Omega$ with the presence of possible holes. From the mathematical point of view the problem can written by

$$\min_{\omega\subset\Omega \text{ open}} \int_\omega F(x,u)\,dx$$
$$\begin{cases} -\Delta u = f \ \text{ in } \omega \\ u = 0 \ \text{ on } \partial\omega. \end{cases} \tag{1.2}$$

In this work we are interested in considering the couple problem where as in (1.1), we look for the optimal distribution of two conductive materials and, similarly to (1.2), we search the set where the diffusion equations is posed. If we consider a constraint on the amounts of the materials used in the mixture, the problem can be formulated as

$$\min_{\omega^\alpha,\omega^\beta} \int_{\omega^\alpha\cup\omega^\beta} F(x,u)\,dx$$
$$\begin{cases} -\mathrm{div}\big((\alpha\chi_{\omega^\alpha} + \beta\chi_{\omega^\beta})\nabla u\big) = f \ \text{ in } \omega^\alpha\cup\omega^\beta \\ u = 0 \ \text{ on } \partial(\omega^\alpha\cup\omega^\beta) \\ \omega^\alpha,\omega^\beta \subset \Omega \ \text{ measurable, } \ \omega^\alpha\cup\omega^\beta \ \text{ open, } \ |\omega^\alpha| \le \kappa^\alpha, \ \ |\omega^\beta| \le \kappa^\beta, \end{cases} \tag{1.3}$$

with κ^α, κ^β two positive constants.

The lack of classical solutions of (1.1) and (1.2) is well-known ([10]). In this work, we obtain a relaxed formulation of (1.3), system of optimality conditions, and we provide a numerical algorithm to solve it. We show some numerical experiments ([6]).

 ISBN 978-84-10135-30-7

2. Statement of the problem and relaxation

We are interested in the optimal design problems of the kind of (1.3) with $\Omega \subset \mathbb{R}^N$ a bounded open set, $f \in H^{-1}(\Omega)$, $\alpha, \beta, \kappa^\alpha, \kappa^\beta$, four positive constants with $\alpha < \beta$, and $F : \Omega \times \mathbb{R} \to \mathbb{R}$ such that

$$F(\cdot, s) \text{ is measurable in } \Omega, \ \forall\, s \in \mathbb{R}, \tag{2.1}$$

$$F(x, \cdot) \text{ is continuous in } \mathbb{R}, \text{ a.e. } x \in \Omega, \tag{2.2}$$

$$\exists\, r \in L^1(\Omega),\ \gamma > 0, \text{ such that } |F(x,s)| \leq r(x) + \gamma |s|^2, \quad \forall\, s \in \mathbb{R}, \text{ a.e. } x \in \Omega. \tag{2.3}$$

Since as we said in the introduction the problem has no solution in general, we look for a relaxed formulation, it will be obtained using the homogenization theory. In this way we will use the following classical result due to S. Spagnolo ([12]). See also [11].

Theorem 2.1 *Let $\Omega \subset \mathbb{R}^N$ be a bounded open set, and $A_n \in L^\infty(\Omega)^{N\times N}$ a sequence of symmetric matrix functions such that there exist $\alpha, \beta > 0$ satisfying*

$$\alpha |\xi|^2 \leq A_n(x)\xi \cdot \xi \leq \beta |\xi|^2, \quad \forall\, \xi \in \mathbb{R}^N, \ \text{a.e. } x \in \Omega. \tag{2.4}$$

Then, for a subsequence of n, still denoted by n, there exists a symmetric matrix function $A \in L^\infty(\Omega)^{N\times N}$, which also satisfies (2.4), such that for every $f \in H^{-1}(\Omega)$, the solution u_n of

$$\begin{cases} -\mathrm{div}(A_n \nabla u_n) = f \ \text{in } \Omega \\ u_n \in H_0^1(\Omega), \end{cases}$$

satisfies

$$u_n \rightharpoonup u \ \text{in } H_0^1(\Omega), \qquad A_n \nabla u_n \rightharpoonup A \nabla u \ \text{in } L^2(\Omega)^N,$$

with u the solution of

$$\begin{cases} -\mathrm{div}(A \nabla u) = f \ \text{in } \Omega \\ u \in H_0^1(\Omega). \end{cases}$$

We say that A_n H-converges to A and we write $A_n \overset{H}{\rightharpoonup} A$.

We are interested in the case where the domains also varies. In this sense it is necessary to recall some results about capacity.

Definition 2.2 *For a bounded open set $\Omega \subset \mathbb{R}^N$ and $E \subset \Omega$, we define the capacity of E in Ω as*

$$\mathrm{Cap}(E, \Omega) := \inf\left\{ \int_\Omega |\nabla \varphi|^2 dx : \ \varphi \in H_0^1(\Omega),\ \varphi \geq 1 \ \text{a.e. in a neighbourhood of } E \right\}.$$

Definition 2.3 *A set $U \subset \Omega$ is said to be quasi-open if for every $\varepsilon > 0$, there exists $G \subset \Omega$ open such that $\mathrm{Cap}(U \Delta G, \Omega) < \varepsilon$. The complementary in Ω of a quasi-open set U is said to be quasi-closed.*

We define $M_0(\Omega)$ as the set of non-negative Borel measures which vanish on the null-capacity sets of Ω and satisfy

$$\mu(E) = \inf\left\{ \mu(U) : \ E \subset U,\ U \text{ quasi-open} \right\}.$$

It is important to remark that the elements of $M_0(\Omega)$ are not necessarily Radon measures. They can take a infinity values in compact subets of Ω. Namely, for every $\mu \in M_0(\Omega)$, there exists a unique quasi-closed set that we will note by C_μ such that

$$\mu = \infty_{C_\mu} \text{ in } C_\mu, \qquad \mu \text{ is } \sigma\text{-finite in } \Omega \setminus C_\mu,$$

where ∞_{C_μ} is the measure in $M_0(\Omega)$ defined as

$$\infty_{C_\mu}(E) = \begin{cases} \infty & \text{if } \mathrm{Cap}(E \cap C_\mu, \Omega) > 0 \\ 0 & \text{if } \mathrm{Cap}(E \cap C_\mu, \Omega) = 0. \end{cases}$$

An extension of Theorem 2.1 for the case where the open set Ω also varies is given by the following theorem due to G. Dal Maso and F. Murat ([4]).

Theorem 2.4 *Assume $\Omega \subset \mathbb{R}^N$ a bounded open set, $A_n \in L^\infty(\Omega)^{N\times N}$ symmetric, which satisfies (2.4) and $\mu_n \in M_0(\Omega)$. Then, for a subsequence of n still denoted by n, there exits a symmetric matrix $A \in L^\infty(\Omega)^{N\times N}$ and a measure $\mu \in M_0(\Omega)$ such that A_n H-converges to A and for every $f \in H^{-1}(\Omega)$ the sequence of solutions of*

$$\begin{cases} -\mathrm{div}(A_n \nabla u_n) + \mu_n u_n = f \ \text{ in } \Omega \\ u_n \in H_0^1(\Omega) \cap L^2_{\mu_n}(\Omega), \end{cases} \tag{2.5}$$

converges weakly in $H_0^1(\Omega)$ to the unique solution of

$$\begin{cases} -\mathrm{div}(A \nabla u) + \mu u = f \ \text{ in } \Omega \\ u \in H_0^1(\Omega) \cap L^2_{\mu}(\Omega). \end{cases} \tag{2.6}$$

We will write

$$(A_n, \mu_n) \overset{H\gamma}{\rightharpoonup} (A, \mu). \tag{2.7}$$

Definition 2.5 *For $p \in [0,1]$, we denote by $m^-(p)$ and $m^+(p)$ the harmonic and arithmetic mean values of α and β with proportions p and $1-p$ respectively, i.e.*

$$m^-(p) = \left(\frac{p}{\alpha} + \frac{1-p}{\beta}\right)^{-1}, \qquad m^+(p) = p\alpha + (1-p)\beta.$$

We also define $K(p)$ as the set of symmetric matrices $M \in \mathbb{R}^{N\times N}$ such that their eigenvalues $\lambda_1 \le \cdots \le \lambda_N$ satisfy

$$\begin{cases} m^-(p) \le \lambda_i \le m^+(p), \ 1 \le i \le N \\ \displaystyle\sum_{i=1}^N \frac{1}{\lambda_i - \alpha} \le \frac{1}{m^-(p) - \alpha} + \frac{N-1}{m^+(p) - \alpha} \\ \displaystyle\sum_{i=1}^N \frac{1}{\beta - \lambda_i} \le \frac{1}{\beta - m^-(p)} + \frac{N-1}{\beta - m^+(p)}. \end{cases}$$

Remark 2.6 *The set $K(p)$ corresponds with the H-closure of two isotropic materials with fixed proportion p and $1-p$, respectively, which was obtained in [13].*

Using Theorem 2.4 we have obtained in [6] the following result adapted to problem (1.3).

Theorem 2.7 *Assume $\Omega \subset \mathbb{R}^N$ a bounded open set, $\mu_n \in M_0(\Omega)$, $\theta_n^\alpha, \theta_n^\beta \in L^\infty(\Omega;[0,1])$, and $A_n \in L^\infty(\Omega \setminus C_{\mu_n})^{N\times N}$ such that*

$$\theta_n^\alpha + \theta_n^\beta \le 1 \ \text{ a.e. in } \Omega, \quad \theta_n^\alpha + \theta_n^\beta = 1 \ \text{ a.e. in } \Omega \setminus C_{\mu_n}, \quad A_n \in K(\theta_n^\alpha) \ \text{ a.e. in } \Omega \setminus C_{\mu_n}. \tag{2.8}$$

Then, there exist a subsequence of n, still denoted by n, $\mu \in M_0(\Omega)$, $\theta^\alpha, \theta^\beta \in L^\infty(\Omega,[0,1])$, and $A \in L^\infty(\Omega \setminus C_\mu)^{N\times N}$, satisfying

$$\theta^\alpha + \theta^\beta \le 1 \ \text{ a.e. in } \Omega, \quad \theta^\alpha + \theta^\beta = 1 \ \text{ a.e. in } \Omega \setminus C_\mu, \quad A \in K(\theta^\alpha) \ \text{ a.e. in } \Omega \setminus C_\mu, \tag{2.9}$$

such that

$$\theta_n^\alpha \overset{*}{\rightharpoonup} \theta^\alpha, \quad \theta_n^\beta \overset{*}{\rightharpoonup} \theta^\beta \qquad \text{in } L^\infty(\Omega), \tag{2.10}$$

and such that for every $f \in H^{-1}(\Omega)$, the sequence of solutions u_n of (2.5) converges weakly in $H_0^1(\Omega)$ to the solution u of (2.6).

From Theorem 2.7 we can obtain the following relaxation version of (1.3).

Theorem 2.8 *Let $\Omega \subset \mathbb{R}^N$ be a bounded open set, $f \in H^{-1}(\Omega)$ and F satisfying 2.1, 2.2 and 2.3. Then a relaxed formulation of (1.3) is given by*

$$\min \int_\Omega F(x,u)dx$$

$$\begin{cases} -\mathrm{div}(A\nabla u) + \mu u = f \ \text{ in } \Omega, \quad u \in H_0^1(\Omega) \cap L^2_\mu(\Omega) \\ \mu \in M_0(\Omega), \quad \theta^\alpha, \theta^\beta \in L^\infty(\Omega;[0,1]), \quad A \in K(\theta^\alpha) \ \text{a.e. in } \Omega \setminus C_\mu \\ \theta^\alpha + \theta^\beta = 1 \ \text{ a.e. in } \Omega \setminus C_\mu, \ \theta^\alpha + \theta^\beta \le 1 \ \text{ in } \Omega, \ \displaystyle\int_\Omega \theta^\alpha dx \le \kappa^\alpha, \ \int_\Omega \theta^\beta \le \kappa^\beta. \end{cases} \tag{2.11}$$

Remark 2.9 *The set $K(\theta^\alpha)$ has an explicit but complex identification, in this sense, having in mind that in the relaxed formulation it is necessary $A\nabla u$ only, we can replace this set by*

$$\mathrm{Sp}(A) \subset \left[m^-(\theta^\alpha), m^+(\theta^\alpha)\right] \text{ a.e. in } \Omega \setminus C_\mu.$$

Then, an alternative formulation of Theorem 2.8 is the following.

$$\min \int_\Omega F(x,u)dx$$

$$\begin{cases} -\mathrm{div}(A\nabla u) + \mu u = f \text{ in } \Omega, \quad u \in H_0^1(\Omega) \cap L^2_\mu(\Omega) \\ \mu \in M_0(\Omega), \quad \theta^\alpha \in L^\infty(\Omega \setminus C_\mu; [0,1]), \quad A \in L^\infty(\Omega \setminus C_\mu)^{N\times N} \text{ symmetric} \\ \mathrm{Sp}(A) \subset [m^-(\theta^\alpha), m^+(\theta^\alpha)] \text{ a.e. in } \Omega \setminus C_\mu, \quad |\Omega \setminus C_\mu| - \kappa^\beta \le \int_{\Omega\setminus C_\mu} \theta^\alpha dx \le \kappa^\alpha. \end{cases} \tag{2.12}$$

3. Numerical Algorithm

We propose a numerical algorithm to solve the relaxed problem (2.12). We have two controls in the problem, the matrix A and the measure μ, since it can take the value $+\infty$, in order to get an approximation let us use a truncation corresponding to take μ as a measurable function taking values in $[0,n]$ with n a positive constant, large enough, we could identify the set C_μ with the set $\{\mu = n\}$. Then,

$$\int_{\Omega\setminus C_\mu} \theta^\alpha dx \le \kappa^\alpha \text{ replaced by } \int_{\{\mu<n\}} \theta^\alpha dx \le \kappa^\alpha \Leftrightarrow \int_\Omega \theta^\alpha \chi_{\{[0,n)\}}(\mu)dx \le \kappa^\alpha.$$

However the function $(s,\mu) \in [0,1]\times[0,\infty) \to s\chi_{\{[0,n)\}}(\mu)$ is not convex. Thus, it is more convenient to use its convex hull given by

$$(s,\mu) \in [0,1]\times[0,\infty) \to \left(s - \frac{\mu}{n}\right)^+.$$

Thus, we replace (2.12) by

$$\min \int_\Omega F(x,u)dx$$

$$\begin{cases} -\mathrm{div}(A\nabla u) + \mu u = f \text{ in } \Omega, \quad u \in H_0^1(\Omega) \\ \mu \in L^\infty(\Omega;[0,n]), \quad \theta \in L^\infty(\Omega;[0,1]), \quad A \in L^\infty(\Omega)^{N\times N} \text{ symmetric} \\ \mathrm{Sp}(A) \subset [m^-(\theta), m^+(\theta)] \text{ a.e. in } \Omega \\ \int_\Omega \left(\theta - \frac{\mu}{n}\right)^+ dx \le \kappa^\alpha, \quad \int_\Omega \left(1 - \theta - \frac{\mu}{n}\right)^+ dx \le \kappa^\beta. \end{cases} \tag{3.1}$$

The following theorem is proved in [6].

Theorem 3.1 *Problem (3.1) has at least one solution for every $n \in \mathbb{N}$. Moreover, for every sequence of solutions (θ_n, A_n, μ_n) of (3.1), there exist a subsequence, still denoted by n, and a solution $(\hat\theta^\alpha, \hat\theta^\beta, \hat A, \hat\mu)$ of (2.11) such that denoting by u_n and $\hat u$ the solutions of the respective state equations, we have*

$$\begin{cases} u_n \rightharpoonup \hat u \text{ in } H_0^1(\Omega), \quad (A_n, \mu_n) \overset{H\gamma}{\rightharpoonup} (\hat A, \hat\mu) \\ \left(\theta_n - \frac{\mu_n}{n}\right)^+ \overset{*}{\rightharpoonup} \hat\theta^\alpha, \quad \left(1 - \theta_n - \frac{\mu_n}{n}\right)^+ \overset{*}{\rightharpoonup} \hat\theta^\beta \quad \text{in } L^\infty(\Omega). \end{cases} \tag{3.2}$$

Moreover

$$\lim_{n\to\infty} \int_\Omega F(x,u_n)\,dx = \int_\Omega F(x,\hat u)\,dx. \tag{3.3}$$

Having in mind the convexity of the set of controls, for a given set of controls (θ_k, A_k, μ_k) we search some new controls

$$\begin{cases} \theta_{k+1} = \theta_k + \varepsilon_k(\hat\theta - \theta_k), \\ A_{k+1} = A_k + \varepsilon_k(\hat A - A_k), \\ \mu_{k+1} = \mu_k + \varepsilon_k(\hat\mu - \mu_k), \end{cases} \tag{3.4}$$

such that the cost function decreases.

We propose to use a gradient descent method where the volume constraints are introduced by Lagrange multipliers (to determine) in the cost functional, these Lagrange multipliers are obtained using the Uzawa method. For more details for the algorithm see [6].

We put u_k the solutin of

$$\begin{cases} -\mathrm{div}(A_k \nabla u_k) + \mu_k u_k = f \ \text{ in } \Omega \\ u_k \in H_0^1(\Omega). \end{cases} \tag{3.5}$$

We introduce the adjoint state p_k as follow:

$$\begin{cases} -\mathrm{div}(A_k \nabla p_k) + \mu_k p_k = \partial_s F(x, u_k) \ \text{ in } \Omega \\ p_k \in H_0^1(\Omega), \end{cases} \tag{3.6}$$

and the functions

$$\begin{cases} E_k^+ = \dfrac{|\nabla u_k||\nabla p_k| + \nabla u_k \cdot \nabla p_k}{2}, \\ \\ E_k^- = \dfrac{|\nabla u_k||\nabla p_k| - \nabla u_k \cdot \nabla p_k}{2}. \end{cases} \tag{3.7}$$

We fix a number $n \in \mathbb{N}$, large enough and note $I_k = \int_\Omega F(x, u_k)\, dx$. The algorithm is the following:

- Initialization: consider $\lambda_{0,1}, \lambda_{0,2} \geq 0$, $\theta_0 \in L^\infty(\Omega; [0,1])$, $A_0 \in L^\infty(\Omega)^{N\times N}$, $\mathrm{Sp}(A_0) \subset [m^-(\theta), m^+(\theta)]$, $\mu_0 \in L^\infty(\Omega; [0,n])$, $\rho > 0$ small and $\bar{j} \in \mathbb{N}$.
- for $k \geq 0$, iterate until convergence as follow:
  - We compute the solutions u_k, p_k of (3.5) and (3.6) respectively, and later E_k^+, E_k^- defined by (3.7).
  - We denote $\lambda_{k,1}^0 = \lambda_{k,1}, \lambda_{k,2}^0 = \lambda_{k,2}$, then for $j \leq \bar{j} - 1$, we define $(\lambda_{k,1}^{j+1}, \lambda_{k,2}^{j+1})$ by

$$\begin{cases} \lambda_{k,1}^{j+1} = \left(\lambda_{k,1}^j + \rho\Big(\displaystyle\int_\Omega \Big(\theta_k^j - \frac{\mu_k^j}{n}\Big)^+ dx - \kappa^\alpha\Big)\right)^+ \\ \\ \lambda_{k,2}^{j+1} = \left(\lambda_{k,2}^j + \rho\Big(\displaystyle\int_\Omega \Big(1 - \theta_k^j - \frac{\mu_k^j}{n}\Big)^+ dx - \kappa^\beta\Big)\right)^+, \end{cases} \tag{3.8}$$

    with θ_k^j, μ_k^j are defined by Proposition 4.2 in [6].
  - We take $\lambda_{k,1} = \lambda_{k,1}^{\bar{j}}, \lambda_{k,2} = \lambda_{k,2}^{\bar{j}}, \hat{\theta} = \theta_k^{\bar{j}}, \hat{\mu} = \mu_k^{\bar{j}}$ and $\hat{A}$ as a symmetric matrix function in $L^\infty(\Omega)^{N\times N}$ such that

$$\begin{cases} \hat{A}\nabla u_k = \dfrac{m^+(\hat{\theta}) + m^-(\hat{\theta})}{2}\nabla u_k + \dfrac{m^+(\hat{\theta}) - m^-(\hat{\theta})}{2}\dfrac{|\nabla u_k|}{|\nabla p_k|}\nabla p_k \quad \text{a.e. in } \{\nabla p_k \neq 0\} \\ \\ \hat{A}\nabla p_k = \dfrac{m^+(\hat{\theta}) + m^-(\hat{\theta})}{2}\nabla p_k + \dfrac{m^+(\hat{\theta}) - m^-(\hat{\theta})}{2}\dfrac{|\nabla p_k|}{|\nabla u_k|}\nabla u_k \quad \text{a.e. in } \{\nabla u_k \neq 0\}. \end{cases} \tag{3.9}$$

    with $\mathrm{Sp}(\hat{A}) \subset [m^-(\hat{\theta}), m^+(\hat{\theta})]$, a.e. in Ω where $m^-(\hat{\theta})$ and $m^+(\hat{\theta})$ the harmonic and arithmetic mean values of α and β with proportions $\hat{\theta}$ and $1 - \hat{\theta}$ respectively.
  - For $\varepsilon_k \in (0,1]$, we update $\theta_{k+1}, A_{k+1}, \mu_{k+1}$ by (3.4).
- Stop if convergence: $\frac{|I_k - I_{k-1}|}{|I_0|} < tol$, for $tol > 0$ small.

We finish this section showing some numerical experiments based in the algorithms described above. The computation has been carried out using the free software FreeFem++ v4.5 ([8], available in http://www.freefem.org). The figures are obtained using Paraview 5.10.1 (available at https://www.kitware.com/open-source/# paraview), which is free also.

We use P_1-Lagrange finite element approximations for u_k and p_k, solutions of the state and costate equations respectively, and P_0-Lagrange finite element approximations for control variables, (θ_k, A_k, μ_k). For all simulations we consider $\Omega = [0,1]^2$, $\alpha = 1$ and $\beta = 2$.

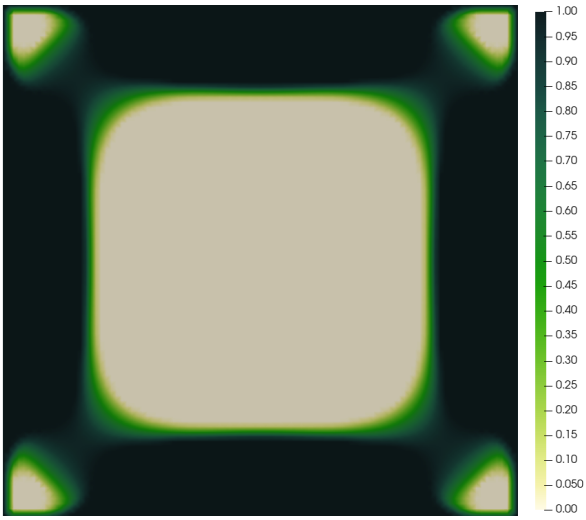


Fig. 1 Example 1: $\kappa^\alpha = \kappa^\beta = \frac{1}{2}$: optimal θ.

Example 1. We consider $F(x,u) = -u$, $f = 1$ and $\kappa^\alpha = \kappa^\beta = 0.5$. This problem has been solved by several authors in the case where we only optimize the matrix A and fixed $\mu \equiv 0$ ([1], [5], [7], [9]). We have considered $n = 10^4$, and we recover the optimal measure $\mu = 0$ and (θ^α, A) given by the previous works, see Figure 1.

Example 2. We consider $F(x,u) = \dfrac{1}{2}\displaystyle\int_\Omega |u-1|^2\,dx$, $f = 1$ and different values of κ^α and κ^β. For the first simulation we consider $\kappa^\alpha = 0.35$ and $\kappa^\beta = 0.3$, in this case there is not enough material to fill out all the domain Ω, thus we expect that the optimal $\mu \not\equiv 0$ defines a smaller domain, see Figure 2.

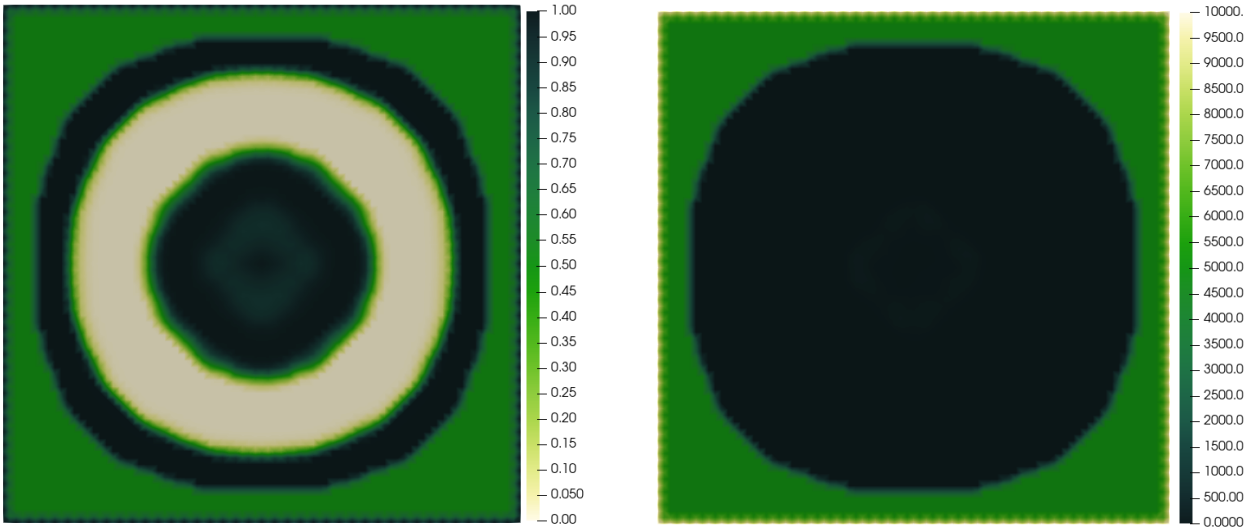


Fig. 2 Example 2, $\kappa^\alpha = 0.35$ and $\kappa^\beta = 0.3$: computed optimal θ (left), computed optimal μ (right).

For a second simulation we consider $\kappa^\alpha = 0.43$ and $\kappa^\beta = 0.62$. In this case, as we expect all the domain is filled out using both materials and holes do not appears, and $\mu \equiv 0$, see Figure 3.

Finally, in Figure 4 we show the convergence of the algorithm for Example 2 in the case $\kappa^\alpha = 0.43$ and $\kappa^\beta = 0.62$. For the rest of the numerical simulations the convergence evolution is similar.

Acknowledgements

The authors were partially supported by Grant PID2020-116809GB-I00 of the Government of Spain.

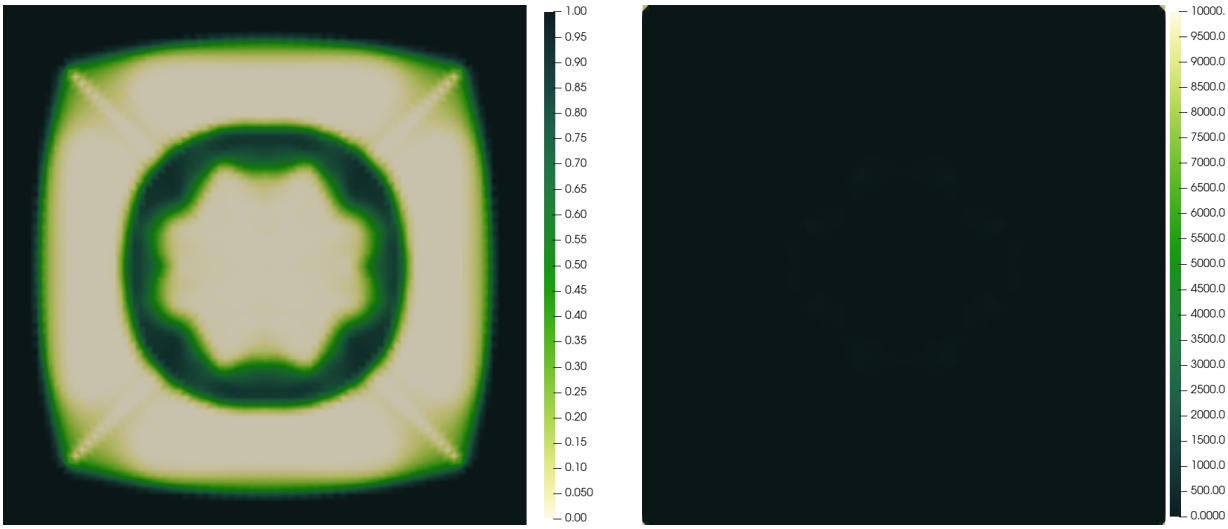


Fig. 3 Example 2, $\kappa^{\alpha} = 0.43$ and $\kappa^{\beta} = 0.62$: computed optimal θ (left), computed optimal μ (right).

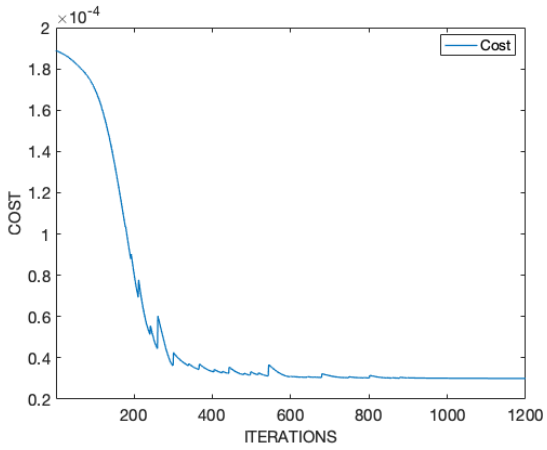


Fig. 4 Example 2, $\kappa^{\alpha} = 0.43$ and $\kappa^{\beta} = 0.62$: cost evolution.

References

[1] G. Allaire. *Shape optimization by the homogenization method.* Appl. Math. Sci. 146, Springer-Verlag, New York, 2002.

[2] M.P. Bendsøe, O. Sigmund. *Topology optimization.* Springer-Verlag, Berlin, 2003.

[3] G. Buttazzo, G. Dal Maso. *Shape optimization for Dirichlet problems: relaxed formulation and optimality conditions.* Appl. Math. Optim. 23 (1991), 17–49.

[4] G. Dal Maso, F. Murat. *Asymptotic behaviour and correctors for linear Dirichlet problems with simultaneously varying operators and domains.* Ann. Inst. H. Poincaré Anal. Non Linéaire 21 (2004), 445–486.

[5] J. Casado-Díaz. *Optimal design of multi-phase materials with a cost functional that depends nonlinearly on the gradient.* Springer-Briefs in Mathematics. Springer, Cham, 2022.

[6] J. Casado-Díaz, M. Luna-Laynez, F. Maestre. *Control problems in the coefficients and the domain for linear elliptic equations.* Submitted.

[7] R. Glowinski. *Numerical simulation for some applied problems originating from continuum mechanics.* In *Trends in applications of pure mathematics to mechanics*. Symp., Palaiseau/France 1983, Lecture Notes in Physics 195 (1984), 96–145.

[8] F. Hecht. *New development in FreeFem++.* J. Numer. Math. 20, 251–265 (2012).

[9] B. Kawohl, J. Stara, G.Wittum. *Analysis and numerical studies of a problem of shape design.* Arc. Rational Mech. Anal. 114, (1991), 343–363.

[10] F. Murat. *Théorèmes de non existence pour des problèmes de contrôle dans les coefficients.* C.R.A.S Sci. Paris A 274 (1972), 395-398.

[11] F. Murat, L. Tartar. *Calcul des variations et homogénéisation*. In *Les méthodes de l'homogénéisation: theorie et applications en physique*, Eirolles, Paris, 1985, 319–369. English translation : F. Murat, L. Tartar. *Calculus of variations and homogenization*. In *Topics in the Mathematical Modelling of Composite Materials*, ed. by L. Cherkaev, R.V. Kohn. Progress in Nonlinear Diff. Equ. and their Appl., 31, Birkaüser, Boston, 1998, 139–174.

[12] S. Spagnolo. *Sulla convergenza di soluzioni di equazioni paraboliche ed ellittiche*. Ann. Scuola Norm. Sup. Pisa Cl. Sci. 22 (1968), 571–597.

[13] L. Tartar. *Estimations fines de coefficients homogénéisés*. In Ennio de Giorgi colloquium (Paris, 1983), (ed. P. Kree). Research Notes in Math. 125, Pitman, London, 1985, 168–187.

[14] L. Tartar. *The general theory of homogenization. A personalized introduction*. Springer, Berlin Heidelberger, 2009.

FRENCH-GERMAN-SPANISH CONFERENCE ON OPTIMIZATION
Gijón, June 18-21, 2024
(pp. 87-93)

Duality for infinite horizon relaxed control problems

Ilya Dikariev[2], **Valeriya Lykina**[3], **Sabine Pickenhain**[1]

1. sabine.pickenhain@b-tu.de Brandenburg Technical University Cottbus-Senftenberg, Germany
2. dikarill@b-tu.de Brandenburg Technical University Cottbus-Senftenberg, Germany
3. Valeriya.Lykina@b-tu.de Brandenburg Technical University Cottbus-Senftenberg, Germany

1. Introduction

We consider control problems $(P)_\infty$ whose objective functional is in an economic context a utility functional,

$$J(x,u) \quad = \quad \int_0^\infty W(x(t),u(t))\, e^{-\varrho t}\, dt \quad \longrightarrow Max\,! \tag{1.1}$$

where W is an instantaneous utility function and ϱ is a positive or zero discount rate.
The objective can also be an energy functional in mechanical or quantum mechanical systems, or it can be chosen in such a way that the asymptotic and exponential controllability of the system is guaranteed,

$$J(x,u) \quad = \quad \int_0^\infty \frac{1}{2}\left(x(t)^T Q(t)x(t) + u(t)^T R(t)u(t)\right) e^{\beta t}\, dt \quad \longrightarrow Min\,! \tag{1.2}$$

where $\beta > 0$ assures together with the choice of suitable state spaces the exponential stability of the solution.

All the target functionals considered have in common that they are given on an a priori infinite horizon and a weight function occurs in the integrand of the objective.

We consider non-linear, non-autonomous dynamical systems. Consequently, one has to expect that convexity assumptions, which are usually required for existence results, are not fulfilled. We take this into account by passing to an optimal control problem with relaxed controls $(\bar{P})_\infty$. Dual-based methods for solving the problems are proposed. It turns out that $(P)_\infty$ and $(\bar{P})_\infty$ have a common dual problem. A Lotka-Volterra model is presented as an application.

2. Problem statement

The following problem $(\bar{P})_\infty$ is considered:

$$\begin{aligned} J(x,\mu) \quad &= \quad \int_0^\infty \int_U r(t,x(t),\mathrm{v})\, d\mu_t(\mathrm{v}) e^{-\varrho t}\, dt \longrightarrow Min\,! \\ x \quad &\in \quad W_2^{1,n}((0,\infty),\nu), \quad \mu \in \mathcal{M}_U,\ U \in comp(\mathbb{R}^m), \\ \dot{x}(t) \quad &= \quad \int_U f(t,x(t),\mathrm{v}) d\mu_t(\mathrm{v}) \text{ a.e. on } (0,\infty),\ x(t_0) = x^0. \end{aligned}$$

All integrals are to be understood in the Lebesgue sense. The control domain U is assumed to be compact. $W_2^{1,n}((0,\infty),\nu)$ is a weighted Sobolev-space and relaxed controls are taken from a family of probability measures $\mathcal{M}_U$, introduced in the next section.

3. Spaces of states and controls

3.1. Control spaces

The relaxed controls μ are taken from a regular family of probability measures $\mathcal{M}_U$, [4].

Definition 3.1 A relaxed control $\{\mu_t\}_{t\in\mathbb{R}_+}$ is a family of probability measures that has the following properties:

1. supp $\mu_t \subseteq U$ a.e. on $\mathbb{R}_+$,
2. μ_t is a probability measure on U a.e. on $\mathbb{R}_+$,

 ISBN 978-84-10135-30-7

3. For all continuous functions with compact support, $g \in C_c(\mathbb{R}_+ \times U)$, the function

$$t \longrightarrow \int_U g(t,v)\, d\mu_t(v)$$

is Lebesgue - measurable.

The motivation for introducing relaxed controls is given by the following arguments:

Remark 3.2
1. In general nonlinear systems cannot be stabilized using a continuous closed loop control $U(x)$, even if each state separately can be driven asymptotically to the origin.
2. Sometimes it can be stabilized with a *continuous closed loop relaxed control*.
3. Relaxed control-type stabilization is used both in theory and in practice; the method is known as *dithering*, see [1].

3.2. State spaces

A weighted Sobolev space $W_2^{1,n}((0,\infty),\nu)$ with a suitable weight function ν is chosen as the state space.
The introduction of the weighted Sobolev space $W_2^{1,n}((0,\infty),\nu)$ is motivated by the following facts. Density and weight functions appear naturally in the objective functionals. If a classical Sobolev space $W_2^{1,n}(0,T)$ is usually used as state space for control problems with bounded time interval $[0,T]$, the limit transition $T \to \infty$ leads in a natural way to an improper integral

$$\lim_{T\to\infty} \int_0^T f(x)\, dx$$

which, in general not coincides with the Lebesgue - integral, see [9].
While in the case of bounded intervals the elements of the Banach space $W_1^1((0,T))$ have a continuous representative and thus the space $W_1^1((0,T))$ can be identified with the space of absolutely continuous functions $AC((0,T))$, the continuation of this space to $AC_{loc}((0,\infty))$ loses the Banach space structure. This is an important theoretical motivation to switch to weighted Sobolev spaces as Banach spaces in the problem definition.

Definition 3.3 (weight function/density function) Let $\mathbb{R}_+ := [0,\infty)$. A continuous function $\nu : \mathbb{R}_+ \to \mathbb{R}_+$ is called *weight function* if ν and $\nu^{-1} \in L_{1,loc}(\mathbb{R}_+)$. If for a weight function ν also holds

$$\int_{\mathbb{R}_+} \nu(t)dt < \infty$$

we call this *density function*. Otherwise we name it *proper weight function*.

Definition 3.4 Let $M^n(\mathbb{R}_+)$ be the set of measurable vector functions on $\mathbb{R}_+$. By means of a weight function ν, we define the *weighted Lebesgue space*

$$L_2^n(\mathbb{R}_+,\nu) \quad = \quad \left\{ x \in M^n(\mathbb{R}_+) \mid \|x\|^2_{L_2^n(\mathbb{R}_+,\nu)} := \int_{\mathbb{R}_+} x^T(t)x(t)\nu(t)dt < \infty \right\} \tag{3.1}$$

the *weighted Sobolev space*

$$W_2^{1,n}(\mathbb{R}_+,\nu) = \left\{ x \in M^n(\mathbb{R}_+) \;\middle|\; x \in L_2^n(\mathbb{R}_+,\nu)\,,\ \mathcal{D}x \in L_2^n(\mathbb{R}_+,\mu) \right\}. \tag{3.2}$$

where $\mathcal{D}x$ denotes the distributional derivative (shortly denoted by x'), see [8], p. 11 ff. With the introduced norm $L_2^n(\mathbb{R}_+,\nu)$ becomes a Hilbert space. With

$$\|x\|_{W_2^{1,n}(\mathbb{R}_+,\nu)} \;=\; \|x\|_{L_2^n(\mathbb{R}_+,\nu)} \;+\; \|\mathcal{D}x\|_{L_2^n(\mathbb{R}_+,\nu)}\,, \tag{3.3}$$

$W_2^{1,n}(\mathbb{R}_+,\nu)$ becomes a Hilbert space as well (this can be confirmed analogously to [8].

The following properties of functions in weighted Sobolev spaces should be mentioned here explicitly.

Remark 3.5 1. Let $x \in W_2^1(\mathbb{R}_+, v)$, $\|x\| \leq K$, $v(t) = e^{\beta t}$, $\beta > 0$, then x is exponentially stable,

$$|x(t)| \leq \left(|x(0)| + C\,K\sqrt{t}\right)e^{-\frac{\beta}{2}t}.$$

2. Let $x \in W_2^1(\mathbb{R}_+, v)$ and $y \in W_2^1(\mathbb{R}_+, v^{-1})$, $v(t) = e^{\beta t}$, $\beta > 0$, then $x\,y$ is asymptotically stable,

$$x\,y \in W_1^1(\mathbb{R}_+) \quad \text{and} \quad \lim_{t\to\infty} x(t)\,y(t) = 0.$$

For the proofs see [11].

4. Optimality notions

In comparison to the literature, see [3], [5], where overtaking or weakly overtaking optimality is mainly used as optimality criterion, the classical comparison of Lebesgue integrals in the objective of $(\bar{P})$ is used here. The admissible domain $\mathcal{A}$ of $(\bar{P})_\infty$ is given by

$$\mathcal{A} := \left\{ (x,\mu) \in W_2^{1,n}(\mathbb{R}_+,\mu) \times \mathcal{M}_U \;\middle|\; \begin{array}{l} x'(t) = \int_U f(t,x(t),\mathrm{v})\mu_t(\mathrm{v})\; a.e.\mathbb{R}_+, \\ \\ x(0) = x^0,\ \mu \in \mathcal{M}_U \end{array} \right\}.$$

Definition 4.1 Let the processes (x,μ), $(x^*,\mu^*) \in \mathcal{A}$ be given. Then the pair (x^*,μ^*) is called *globally optimal in the sense of* **criterion L**, if $J(x^*,\mu^*) < \infty$ and for any pair $(x,\mu) \in \mathcal{A}$ we have

$$J(x^*,\mu^*) \leq J(x,\mu).$$

Under conditions that ensure the existence of the solution, cf. also the contribution by I. Dikariev at the FGS-Conference On Optimization, Gijon, Spain, (2024) , entitled

Existence Theorem for Relaxed Control Problems on Infinite Time Horizon Utilizing Weight Functions

we treat the problem $(\bar{P})$ with dual methods. Here, we mainly refer to the ideas of Carathéodory and Klötzler for the construction of a dual problem. This dual based approach has already been used for special optimal control problems with infinite horizon in [10, 13].

5. Duality

We use a very general scheme for the construction of a dual problem, which goes back to Klötzler, [6]:

Definition 5.1 Let real functionals $F : X \to \bar{\mathbb{R}} := \mathbb{R} \cup +\infty$ and $G : Y \to \bar{\mathbb{R}}$ with arbitrary sets X and Y be given. The problem

$$(\mathbf{D}) \quad G(y) \to \sup! \quad w.r.t.\ y \in Y$$

is called *dual program* to the primary program

$$(\mathbf{P}) \quad F(x) \to \inf! \quad w.r.t.\ x \in X,$$

if the inequality

$$G(y) \leq F(x) \;\; \forall\, x \in X,\ \forall\, y \in Y$$

or equivalently

$$\sup_{y\in Y} G(y) \leq \inf_{x\in X} F(x) \tag{5.1}$$

holds true. Relation (5.1) is called *weak duality relation*. If even the equality holds in (5.1), we say that *the strong duality relation* holds between both problems.

The construction is carried out in the following steps:

Step 1: Partition of the admissible set $\mathcal{A} = X_0 \cap X_1$
Step 2: Define a set Y and a real functional $\Phi(\cdot,\cdot) : X_0 \times Y \to \bar{\mathbb{R}}_+$ with

$$\begin{array}{rcll} \inf\limits_{(x,\mu)\in\mathcal{A}} J(x,\mu) & = & \inf\limits_{(x,\mu)\in X_0} \sup\limits_{S\in Y} \Phi((x,\mu),S) & \text{(equivalence relation)} \\ & \geq & \sup\limits_{S\in Y} \inf\limits_{(x,\mu)\in X_0} \Phi((x,\mu),S) & \end{array}$$

Step 3: For a fixed element $S \in Y$ one sets

$$G(S) := \inf_{(x,\mu)\in X_0} \Phi((x,\mu),S).$$

We realize the scheme and construct a dual Program for $(\bar{\mathbf{P}})_\infty$, with $\nu(t) = e^{\beta t}, \beta > 0$

Step 1: Partition of the admissible set $\mathcal{A} = X_0 \cap X_1$

$$X_0 := \left\{(x,\mu) \in W_2^{1,n}(\mathbb{R}_+,\nu) \times \mathcal{M}_U \,|\, x(0) - x^0 = 0, \mu \in \mathcal{M}_U\right\}$$
$$X_1 := \left\{(x,\mu) \in W_2^{1,n}(\mathbb{R}_+,\nu) \times \mathcal{M}_U \,|\, x'(t) - \textstyle\int_U f(t),x(t),\mathrm{v})d\mu_t(\mathrm{v}) = 0 \text{ a.e. on } (0,\infty)\right\}$$

Step 2: One possible choice for Φ is a Lagrange - functional

$$\Phi_1((x,\mu),S) \quad = \quad J(x,\mu) + \left\langle \underbrace{x'(\cdot) - \textstyle\int_U f(t,x(\cdot),v)d\mu_t(v)}_{\in L_2^n((0,\infty),e^{\beta t})}, \underbrace{\nabla_\xi S(\cdot,x(\cdot))}_{\in L_2^n((0,\infty),e^{-\beta t})} \right\rangle$$

where$\langle\cdot\,,\,\cdot\rangle$ is the scalar product in $L_2^n(\mathbb{R}_+)$, which satisfies

$$\langle \zeta\,, p\,\rangle \le \|\zeta\|_{L_2^n((0,\infty),e^{\beta t})} \|p\|_{L_2^n((0,\infty),e^{-\beta t})}.$$

Then we define the set Y by the following setting:

$$S \in Y \Leftrightarrow S(t,\xi) = y^T(t)\xi \quad \text{and} \quad y \in L_2^n(\mathbb{R}_+,\nu^{-1}), \tag{5.2}$$

$$\Phi_1((x,\mu),y) = J(x,\mu) + \left\langle x'(\cdot) - \int_U f(\cdot,x(\cdot),\mathrm{v})d\mu_t(\mathrm{v}), y(\cdot)\right\rangle_{L_2^n(\mathbb{R}_+)} \tag{5.3}$$

Step 3: Formulation of a dual program (integated version):

$$(\mathbf{D_1}): \qquad G(y) := \inf_{(x,\mu)\in X_0} \Phi_1((x,\mu),y) \to \max! \text{ w.r.t. } y \in L_2^n(\mathbb{R}_+,\nu^{-1}).$$

We can identify the idea of choosing a suitable functional Φ from Carathéodory's approach as well, see [2,3].

It consists of adding an invariant integral to the integral in the objective. Invariance means that the added integral depends on the values of the function S on the boundary of $[0,\infty)$, i.e. on $S(0,x^0)$, only. More precisely, by choosing the function space Y it must be ensured that

$$\int_0^\infty [\int_U r(t,x(t),\mathrm{v})\, d\mu_t(\mathrm{v})\, e^{\beta t} \quad - \quad \frac{d}{dt}S(t,x(t))]\, dt \;= J(x,\mu) + S(0,x^0)$$
$$= \int_0^\infty [\int_U r(t,x(t),\mathrm{v})\, d\mu_t(\mathrm{v})\, e^{\beta t} \quad - \quad [\nabla_\xi^T S(t,x(t)) \int_U f(t,x(t),\mathrm{v})d\mu_t(\mathrm{v}) + S_t(t,x(t))]dt.$$

for all $(x,\mu) \in \mathcal{A}$. Then we conclude

$$J(x,\mu) + S(0,x_0) \quad \ge \quad -\int_0^\infty [\mathcal{H}(t,x(t),\nabla_\xi^T S(t,x(t))) + S_t(t,x(t))]\, dt$$

with the Hamiltonian function

$$\mathcal{H}(t,\xi,\nabla_\xi S(t,\xi)) \quad = \quad \sup\{H(t,\xi,\mathrm{v},S_\xi(t,\xi)) \,|\, \mathrm{v} \in U\}, \tag{5.4}$$

and

$$H(t,\xi,\mathrm{v},S_\xi(t,\xi)) = -r(t,\xi,\mathrm{v})e^{\beta t} + \nabla_\xi^T S(t,\xi) f(t,\xi,\mathrm{v}).$$

This leads together with defect function

$$\Lambda_S(t,\xi) := \mathcal{H}(t,\xi,\nabla_\xi S(t,\xi)) + S_t(t,\xi)$$

in the Hamilton- Jacobi equation and

$$\Lambda_S(t, x^*(t)) = 0 \text{ on } [0, \infty)$$

to the following variant of the dual problem (pointwise version):

$$\begin{array}{rrcl} (\mathbf{D_2}) & \mathbf{G_2(S)} & = & -S(0,x^0) \longrightarrow \max! \\ \text{with respect to} & S & \in & Y \\ & \Lambda_S(t,\xi) & \leq & 0 \quad \forall t \in [0,\infty),\ \forall \xi \\ & \Lambda_S(t,x^*(t)) & = & 0 \quad \forall t \in [0,\infty). \end{array}$$

Remark 5.2 The Hamiltonian for $(P)_\infty$ and $(\bar{P})_\infty$ coincide, since

$$\begin{aligned} \mathcal{H}(t,\xi,\nabla_\xi S(t,\xi)) &= \max\{H(t,\xi,v,S_\xi(t,\xi)) \mid v \in U\} \\ &= \max\{\int_U H(t,\xi,\mathrm{v},S_\xi(t,\xi))d\mu_t(\mathrm{v}) \mid \mu_t \in P_U\} \end{aligned}$$

where P_U is the set of probability measure concentrated on U, see [4]. We conclude that both problems, $(\bar{P})_\infty$ and $(P)_\infty$, have a same dual problem $(\mathbf{D_2})$,

$$\sup((\mathbf{D_2})) \leq \inf((\bar{\mathbf{P}})_\infty) \leq \inf((\mathbf{P})_\infty).$$

6. Applications

The uncontrolled bilinear Lotka-Volterra model considered is

$$\begin{aligned} x_1'(t) &= x_1(t)\,[a - bx_2(t)] \\ x_2'(t) &= x_2(t)\,[-c + dx_1(t)]\,. \end{aligned}$$

6.1. A linearized Lotka-Volterra model

First we transform the non-trivial equilibrium $(\frac{c}{d}, \frac{a}{b}) = (\bar{x}_1, \bar{x}_2)$ of the uncontrolled equilibrium to the zero point. Then we linearize the system around the uncontrolled steady state and look for a bounded control (u_1, u_2) which stabilizes the system exponentially. We arrive at a problem of the following type.

$$(\mathbf{Q}): \quad J(x,u) = \int_0^\infty \frac{1}{2}\left\{(x^T(t)x(t) + u^T(t)u(t))\right\} e^{\beta t}dt \longrightarrow \min!$$

with respect to

$$\begin{array}{rcl} (x,u) & \in & W_2^{1,2}(\mathbb{R}_+, e^{\beta t}) \times L_2^2(\mathbb{R}_+, e^{\beta t}),\ \beta > 0 \\ x'(t) & = & Ax(t) + u(t) \text{ a. e. on } \mathbb{R}_+\,,\ x(0) = x^0\,, \\ u(t) & \in & U := [-1,1] \times [-1,1] \text{ a. e. on } \mathbb{R}_+\,. \end{array}$$

For the detailed assumptions and settings see [7,13]. The corresponding dual problem $(\mathbf{D_Q})$ (integrated version) is

$$(\mathbf{D_Q}): \quad G(y) := -\int_0^\infty \left\{\frac{1}{2}\left[y'(t) + A^T y(t)\right]^T \left[y'(t) + A^T y(t)\right] + \theta(t, y(t))\right\} e^{-\beta t}dt - x_0^T y(0) \longrightarrow \max!$$

w. r. t.

$$y \in W_2^{2,2}(\mathbb{R}_+, e^{-\beta t}) \quad \text{with} \quad x^0 = y'(0) + A^T y(0),$$

with

$$\theta(t, y(t)) = \sum_{i=1}^{2} -\frac{1}{2}\sigma_i^2(t, y(t)) + \sigma_i(t, y(t))y_i(t) \quad \text{and}$$

$$\sigma_i(t, y(t)) = \min\left\{\max\left\{-1, y_i(t)e^{-\beta t}\right\}, 1\right\} e^{\beta t}$$

Remark 6.1 1. In the general construction of the dual problem, (5.2), i.e. $S(t,\xi) = y^T(t)\xi$ and $y \in W_2^{2,2}(\mathbb{R}_+, \nu^{-1})$ is used.

2. The duality construction is carried out with the Lagrange functional (5.3).
3. It can be shown that the Hamilton function (5.4) is smooth.
4. In the dual problem, the inverse weight function appears in the objective functional as well as in the weighted Sobolev space.
5. $(\mathbf{D_Q})$ has an optimal solution.
6. Spectral methods can be applied to approximate the solution.

6.2. A controlled bi-linear Lotka-Volterra model

We transform the steady state of the uncontrolled equilibrium $(\frac{c}{d}, \frac{a}{b}) = (\bar{x}_1, \bar{x}_2)$ to the zero point and look for a bounded control (u_1, u_2) which stabilizes the non-linear system exponentially. We arrive at the following problem:

$$(\tilde{\mathbf{Q}}): \quad J(\tilde{x},u) \;=\; \int_0^\infty \frac{1}{2}\left(\tilde{x}(t)^T Q(t)\tilde{x}(t) + u(t)^T R(t)u(t)\right) e^{\beta t}\,dt \quad \longrightarrow Min!$$

w.r.t.

$$\left(\tilde{x},u\right) \in W_2^{1,2}(\mathbb{R}_+, e^{\beta t}) \times L_2^2(\mathbb{R}_+, e^{\beta t}), \quad \beta > 0$$

with

$$\begin{aligned}
\tilde{x}_1'(t) &= \left[\tilde{x}_1(t) + \frac{c}{d}\right][-b\tilde{x}_2(t) - u_1(t)] \quad \text{a.e. on } \mathbb{R}_+,\\
\tilde{x}_2'(t) &= \left[\tilde{x}_2(t) + \frac{a}{b}\right][d\tilde{x}_1(t) - u_2(t)] \quad \text{a.e. on } \mathbb{R}_+,\\
\tilde{x}_1(0) &= x_1^0 - \frac{c}{d}, \quad \tilde{x}_2(0) = x_2^0 - \frac{a}{b}
\end{aligned}$$

For the duality construction we now use a nonlinear ansatz for S,

$$S(t,\xi) := y_1(t)\ln\left(\xi_1 + \frac{c}{d}\right) + y_2(t)\ln\left(\xi_2 + \frac{a}{b}\right), \quad y \in W_2^{1,2}(\mathbb{R}_+, e^{-\beta t}) \tag{6.1}$$

Then

$$\begin{aligned}
\Phi_2((\tilde{x},u),S) = J(x,u) \;&+\; \int_0^\infty \left(\tilde{x}_1'(t) - \left[\tilde{x}_1(t) + \frac{c}{d}\right][-b\tilde{x}_2(t) - u_1(t)]\right) S_{\xi_1}(t,\tilde{x}(t))\,dt\\
&+\; \int_0^\infty \left(\tilde{x}_2'(t) - \left[\tilde{x}_2(t) + \frac{a}{b}\right][d\tilde{x}_1(t) - u_2(t)]\right) S_{\xi_2}(t,\tilde{x}(t))\,dt\\
= J(\tilde{x},u) \;&+\; \int_0^\infty \left((\ln(\tilde{x}_1 + \frac{c}{d}))'(t) - [-b\tilde{x}_2(t) - u_1(t)]\right) y_1(t)\,dt\\
&+\; \int_0^\infty \left((\ln(\tilde{x}_2 + \frac{a}{b}))'(t) - [d\tilde{x}_1(t) - u_2(t)]\right) y_2(t)\,dt
\end{aligned} \tag{6.2}$$

is well defined and all integrals exist. The final construction of the dual problem in integrated form is similar to that introduced in [7, 11] and [13].

Remark 6.2 1. In the general construction of the dual problem, the nonlinear ansatz of S, (6.1), is used.

2. The duality construction is carried out with the Lagrange functional Φ_2 in (6.2).
3. In the dual problem, the inverse weight function ν^{-1} appears in the objective functional as well as in the weighted Sobolev space.
4. Similar to [13] spectral methods can be applied to approximate the solution of the dual problem.

Acknowledgements

The authors are partially supported by the German Research Foundation grants PI209/8-3 and LY149/2-3.

References

[1] Artstein, Z. Stabilization with relaxed controls *Nonlinear Analysis, Theory Methods and Applications. Vol. 7. No. 11. pp. 1163-1173,(1983).*

[2] Carathéodory, C. Vorlesungen über reelle Funktionen *American Mathematical Society, AMA/Chelsea Series (2004).*

[3] Carlson, D. A. A Caratheodory-Hamilton-Jacobi Theory for Infinite-Horizon Optimal Control Problems *JOTA: VoL 48, No. 2, (1986).*

[4] Gamkrelidze, R.V. Principles of Optimal Control Theory *Plenum Press, New York and London, (1978).*

[5] Haurie A. Existence and Global Asymptotic Stability of Optimal Trajectories for a Class of Infinite-Horizon, Non-convex Systems *JOTA: Vol, 31, No, 4, (1980).*

[6] Klötzler, R. On a general conception of duality in optimal control *Equadiff IV. Proceedings of the Czechoslovak Conference on Differential Equations and their Applications held in Prague, (1977). Springer, Berlin (Lecture Notes in Mathematics 703).*

[7] Kolo, K. Ein dual-basierter Ansatz zur Behandlung linear-quadratischer Optimalsteuerungsprobleme mit einem a priori gegebenen unendlichen Zeithorizont *Dissertationsschrift, BTU Cottbus-Senftenberg, (2021).*

[8] Kufner, A. Weighted Sobolev spaces *Teubner Verlagsgesellschaft, Teubner-Texte zur Mathematik, (1981).*

[9] Lykina, V., Pickenhain S. and Wagner M. Different interpretations of the improper integral objective in an infinite horizon control problem *Journal of Mathematical Analysis and Applications, Vol.340, (2008).*

[10] Lykina, V. Beiträge zur Theorie der Optimalsteuerungsprobleme mit unendlichem Zeithorizont. Dissertation, BTU Cottbus, Germany, (2010).

[11] Pickenhain S. and Burtchen A. Regulator Problems on Unbounded Domains: Stationarity - Optimal Control - Asymptotic Controllability *Vietnam Journal of Mathematics, Vol.46 (2018).*

[12] Pickenhain, S. and Burtchen A. Problems in the Calculus of Variations on Unbounded Intervals, Fourier-Laguerre Analysis and Approximations *Vietnam Journal of Mathematics, (2019).*

[13] Pickenhain, S., Lykina, V. and K. Kolo. Exponential Stabilization of linear control systems with control constraints - a dual based approach *submitted to Optimization, (2022).*

French-German-Spanish Conference on Optimization
Gijón, June 18-21, 2024
(pp. 94-101)

Maximal L^p-regularity of abstract evolution equations modeling closed-loop, boundary feedback control dynamics

Irena Lasiecka[1], **Buddhika Priyasad**[2], **Roberto Triggiani**[3]

1. lasiecka@memphis.edu Department of Mathematical Sciences, University of Memphis, Memphis, TN 38152 USA
1. IBS, Polish Academy of Sciences, Warsaw, Poland
2. priyasad@uni-konstanz.de Department of Mathematics and Statistics, University of Konstanz, Konstanz, Germany
3. rtrggani@memphis.edu Department of Mathematical Sciences, University of Memphis, Memphis, TN 38152 USA

Abstract

We provide maximal L^p-regularity up to the level $T < \infty$ or $T = \infty$ of an abstract evolution equation in Banach space, which captures boundary closed-loop parabolic systems, defined on a bounded multidimensional domain, with finitely many boundary control vectors and finitely many boundary sensors/actuators. Illustrations given include classical parabolic equations as well as Navier-Stokes equations in $L^p(\Omega)$ or $L^q_\sigma(\Omega)$, respectively.

1. The case of boundary controls and boundary sensors/observers, [LPT.6]

Overview

The topic of maximal L^p-regularity was (apparently) first studied in the fundamental paper [Sim] (in Italian) published in 1964. In it, the author considers the generator A of a s.c. (C_0) semigroup e^{At} on the Hilbert space H and shows a definitive result in this setting: that e^{At} possesses maximal L^p-regularity up to T on the Hilbert space H if and only if it is analytic (holomorphic); with $T = \infty$ in case e^{At} is, moreover, (exponentially) uniformly stable. The sophisticated, technical proof was based (as stated in the paper's title) on the theory of singular integrals. This was truly a pioneering paper that stimulated an intense subsequent research activity, both at the abstract Banach space setting as well as at the L^p or Hölder spaces settings for the class of (parabolic) equations. At the general Banach space setting, it was established that maximal L^p-regularity of the s.c. semigroup e^{At} implies that e^{At} is analytic, but not conversely. To date known counterexamples exist in abstract Banach spaces setting, see [HNVW.2, Section 17.4.c]. Instead, the PDE-framework includes: either dynamics defined on the entire multidimensional space; or on half-spaces; or on domains exterior to multidimensional bounded domains; or else on a multidimensional domain Ω, with possibly, open-loop inhomogeneous boundary terms on $\partial\Omega$ in Triebel-Lizorkin spaces, see [DHP]. Similar results are available for Pseudodifferential setting as well. The list of significant papers will likely exceed the length permitted for this extended abstract. Thus, we must constrain ourselves to quote only a few. In contrast, the emphasis of the present extended abstract is quite different. While the setting is still at the abstract Banach space level, the modeled dynamics intend to capture closed-loop boundary feedback (parabolic) problems, with either (i) finitely many boundary controls and interior sensors/actuators [LPT.5]; or else (ii) with finitely many boundary controls and boundary sensors/actuators [LPT.6]. The assumption imposed on the two abstract models are automatically satisfied by the intended, motivating applications. These include, in addition to classical parabolic dynamics, physical important dynamics such as Navier-Stokes equations (particularly in dimension $d = 3$), Boussinesq systems, Magnetohydrodynamics (MHD) systems, etc [LPT.1, LPT.2, LPT.3, LPT.4, LPT.5, LPT.7]. Here maximal L^p-regularity is first established in the Banach space $L^q(\Omega)$, $1 < q < \infty$, or even a suitable Besov space $B^{2-2/p}_{q,p}(\Omega)$ which does not recognize boundary conditions ($1 < p < \frac{2q}{2q-1}$, $q >$ dimension d). Next, such maximal L^p regularity is exploited to obtain (well-posedness as a nonlinear semigroup and) uniform stabilization of the full nonlinear feedback model in the vicinity of an unstable equilibrium solution.

1.1. Abstract setting

The focus of the present section is the operator

$$\begin{cases} A_F = -A(I - GF) : Y \supset \mathcal{D}(A_F) \longrightarrow Y & \text{(1.1a)} \\ \mathcal{D}(A_F) = \{x \in Y : (I - GF)x \in \mathcal{D}(A)\}. & \text{(1.1b)} \end{cases}$$

 ISBN 978-84-10135-30-7

and corresponding abstract equation

$$y_t = A_F y = -A(I - GF)y \tag{1.2}$$

under the following standing assumptions:

(**H.1**) Y is a reflexive Banach space.

(**H.2**) $-A : Y \supset \mathcal{D}(A) \longrightarrow Y$ is the maximal dissipative generator of a C_0-contraction semigroup e^{-At} on Y, $t \geq 0$, which possesses the maximal $L^p(0,T;Y)$-regularity property up to T, either $0 < T < \infty$; or else $T = \infty$, $1 < p < \infty$; in symbols [Dore.1]

$$-A \in MReg\,(L^p(0,T;Y))\,, \quad \text{either } 0 < T < \infty; \text{ or else } T = \infty, 1 < p < \infty;$$

so that, a fortiori, the strongly continuous (s.c.) semigroup e^{At} is analytic (holomorphic) on Y. At the price (harmless for the present note) of replacing A with a suitable translation to the right ($A_k = A + k^2 I$), the fractional powers A^θ, $0 < \theta < 1$, of A are well-defined [Pazy].

(**H.3**) U is another Banach space and G is the ("Green") linear operator satisfying

$$G : \text{continuous } U \longrightarrow \mathcal{D}(A^{\alpha_0}) \subset Y, \text{ or } A^{\alpha_0} G \in \mathcal{L}(U;Y) \tag{1.3}$$

for some $0 < \alpha_0 < 1$.

(**H.4**) F is a linear ("feedback") operator of the form

$$Fz = \langle \gamma z, w \rangle_U \, g, \quad w, g \in U \tag{1.4}$$

where γ is a linear (trace) operator

$$\gamma : \text{continuous } \mathcal{D}(A^\sigma) \subset Y \longrightarrow U, \quad 0 < \sigma < \alpha_0 < 1 \tag{1.5}$$

so that

$$F : \text{continuous } \mathcal{D}(A^\sigma) \subset Y \longrightarrow U. \tag{1.6}$$

$\left[\text{In the applications we shall take } Fz = \sum_{k=0}^{K} \langle \gamma z, w_k \rangle_U \, g_k, \quad w_k, g_k \in U\right]$

Remark 1.1 F is thus unbounded as an operator on Y. For the similar problem considered in [LPT.5] in JDE, F was a bounded operator on Y. The purpose of this work is to extend to the operator (1.1) the result on maximal $L^p(0,T;Y)$-regularity of [LPT.5], $T \leq \infty$. The proof of [LPT.5] requires $F \in \mathcal{L}(Y;U)$. Thus, the proof of the present note is quite different from that in [LPT.5]. See [Las] for abstract parabolic boundary problems.

With reference to assumption (**H.3**) centered on the constant $0 < \alpha_0 < 1$, we introduce two Banach spaces, where $0 \in \rho(A)$,

$$\mathcal{E} \equiv \mathcal{D}(A^{\alpha_0}), \text{ with norm } \|x\|_{\mathcal{E}} \equiv \|x\|_{\mathcal{D}(A^{\alpha_0})} \equiv \|A^{\alpha_0} x\|_Y\,, \tag{1.7}$$

$$E \equiv \left[\mathcal{D}(A^{*(1-\alpha_0)})\right]' \text{ with norm } \|z\|_E \equiv \|z\|_{\left[\mathcal{D}(A^{*(1-\alpha_0)})\right]'} = \left\|A^{-(1-\alpha_0)} z\right\|_Y. \tag{1.8}$$

Accordingly we introduce the following holomorphic interpolation spaces

$$[\mathcal{E}, E]_\theta \equiv \left[\mathcal{D}\left(A^{\alpha_0}\right), \left[\mathcal{D}\left(A^{*(1-\alpha_0)}\right)\right]'\right]_\theta = \begin{cases} \mathcal{D}\left(A^{\alpha_0 - \theta}\right), \; 0 \leq \theta \leq \alpha_0, & \text{(1.9a)} \\ \left[\mathcal{D}\left(A^{*(\theta-\alpha_0)}\right)\right]', \; \alpha_0 \leq \theta \leq 1. & \text{(1.9b)} \end{cases}$$

since $\alpha_0(1-\theta) - (1-\alpha_0)\theta = \alpha_0 - \theta$, with corresponding norm

$$\|x\|_{[\mathcal{E}, E]_\theta} = \|x\|_{\mathcal{D}(A^{\alpha_0 - \theta})} = \left\|A^{\alpha_0 - \theta} x\right\|_Y, \; 0 \leq \theta \leq \alpha_0, \tag{1.10}$$

$$\|z\|_{[\mathcal{E}, E]_\theta} = \|z\|_{\left[\mathcal{D}(A^{*(\theta-\alpha_0)})\right]'} = \left\|A^{-(\theta-\alpha_0)} z\right\|_Y, \; \alpha_0 \leq \theta \leq 1. \tag{1.11}$$

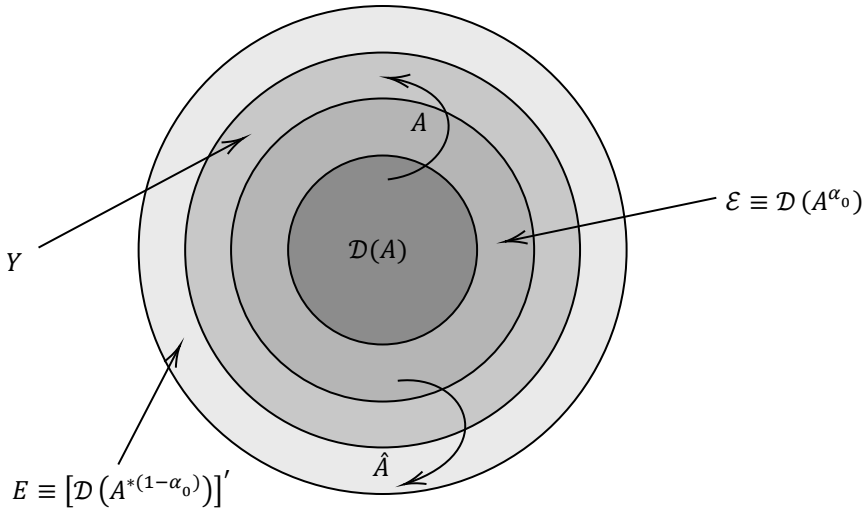


Fig 1: Symbolic illustration of the spaces and operators involved.

1.2. Main Result

Theorem 1.2 *(a) Let $0 < T < \infty$. The operator A_F in (1.1) defined on Y generates a s.c. semigroup $T_F(t)$, which is analytic on Y and, moreover, possesses the maximal $L^p(0,T;Y)$-regularity on Y, $1 < p < \infty$, $T < \infty$: the map*

$$f \to (Lf)(t) = \int_0^t e^{A_F(t-s)} f(s)ds \ \text{ continuous } \ L^p(0,T;Y) \to L^p(0,T;\mathcal{D}(A_F));$$

in symbols, [Dore.1]

$$A_F \in MReg\left(L^p(0,T;Y)\right), \quad 1 < p < \infty, \ T < \infty. \tag{1.12a}$$

(b) Let $T = \infty$. Assume further that the s.c. analytic semigroup $T_F(t)$ is uniformly stable on Y: there exist constants $M \geq 1$, $\delta > 0$, such that

$$\left\|T_F(t)\right\|_{\mathcal{L}(Y)} \leq Me^{-\delta t}, \ t \geq 0. \tag{1.12b}$$

Then, $T_F(t)$ possesses the maximal $L^p(0,\infty;Y)$-regularity on Y, $1 < p < \infty$, $T = \infty$; in symbols [Dore.1]

$$A_F \in MReg\left(L^p(0,\infty;Y)\right), \quad 1 < p < \infty, \ T = \infty. \tag{1.12c}$$

Actually, in the each case (a) and (b), $T_F(t)$ extends/restricts with the same properties - as s.c. analytic, uniformly stable (case (b)) semigroup, with maximal L^p-regularity ($0 < T < \infty$ in case (a), $T = \infty$ in case (b)) - on the space E in (1.8), on the space $\mathcal{E}$ in (1.7), as well as on all holomorphic interpolation spaces (1.9)-(1.11).

The proof of the present Theorem 1.2 with F unbounded as in (1.6), $F \in \mathcal{L}(\mathcal{D}(A^\sigma), U)$ given in [LPT.6], is completely different from the one in [LPT.5]. It is inspired by a proof in [LT.2] about analyticity of a specific parabolic semigroup in an Hilbert setting. It consists of three steps, (i) first, showing L^p-maximal regularity in the larger space E in (1.8); next, (ii) showing L^p-maximal regularity in the smaller space $\mathcal{E}$ in (1.7); and finally, (iii) showing L^p-maximal regularity on Y by interpolation.

In contrast, the proof of [LPT.5] for $F \in \mathcal{L}(Y;U)$ was based on considering A_F^* rather than A_F. With $F \in \mathcal{L}(Y;U)$ and G satisfying $A^\gamma G \in \mathcal{L}(U;Y)$ for some γ, $0 < \gamma < 1$, the expression of A_F makes such form not directly suitable for deducing its maximal regularity on Y, as it would leave the power $A^{1-\gamma}$ on the LHS unaccounted for on Y. The form of A_F^* in [LPT.5] is more amenable to show $A_F^* \in MReg(L^p(0,T;Y^*))$ by perturbation [Dore.2, Theorem 6.2, p311], [KW.1, Remark 1, p426, for $\beta = 1$], [Weis]. Next, to show that the original A_F satisfies $A_F \in MReg(L^p(0,T;Y))$ as desired, paper [LPT.5] employs the result that on the UMD space Y, the property that $A_F \in MReg(L^p(0,T;Y))$ is equivalent to the property that the family, $\tau \in \mathcal{L}(Y)$,

$$\tau \equiv \{tR(it, A_F), \ t \in \mathbb{R}\backslash\{0\}\} \text{ be } R - \text{bounded},$$

[KW.2] where $R(\cdot, A_F)$ denotes the resolvent of A_F. And in the UMD-setting for Y, the R-boundedness property for the family τ is equivalent to the property that the corresponding dual family τ' in $\mathcal{L}(Y^*)$

$$\tau' \equiv \left\{tR(it, A_F^*), \ t \in \mathbb{R}\backslash\{0\}\right\} \text{ be } R - \text{bounded},$$

[HNVW.1, Proposition 8.4.1, p211].

2. Illustrations

For simplicity and brevity of exposition, Example # 1 (for $T < \infty$ and $T = \infty$) will be restricted to a canonical case. More general results can be given by referring to [LT.3, CV, DaV, DaG, Ves].

2.1. Case $0 < T < \infty$.

Example # 1 **The PDE model**: Let Ω be a bounded domain in $\mathbb{R}^d$, $d \geq 2$, with boundary $\partial\Omega \equiv \Gamma$, assumed to be $(d-1)$-dimensional variety with Ω locally on one side of Γ, and sufficiently smooth. We consider the following canonical locally fully boundary closed loop parabolic system on Ω, with boundary control in the Neumann BC and boundary sensing (observations):

$$\begin{cases} \dfrac{\partial y(t,x)}{\partial t} = (\Delta - I)y(t,x) & \text{in } (0,T]\times\Omega \quad (2.1a) \\ y(0,x) = y_0(x) & \text{in } \Omega \quad (2.1b) \\ \dfrac{\partial y(t,\xi)}{\partial \nu} = f(t,\xi) \equiv \displaystyle\sum_{k=0}^{K} (\gamma y(t,\cdot), w_k(\cdot))_\Gamma g_k(\xi) & \quad (2.1c) \\ \qquad\quad \equiv Fy(t,\cdot) & \text{on } (0,T]\times\Gamma \quad (2.1d) \end{cases}$$

(a) Let

$$Y \equiv L^q(\Omega),\ 1 < q < \infty,\ A = -\Delta + I;\ Y \supset \mathcal{D}(A) \to Y \quad (2.2a)$$

$$\mathcal{D}(A) = \left\{ \varphi \in W^{2,q}(\Omega) :\ \left.\frac{\partial \varphi}{\partial \nu}\right|_\Gamma = 0 \right\}. \quad (2.2b)$$

Then $-A$ generates a s.c. contraction, analytic semigroup e^{-At}, $t \geq 0$ on $Y \equiv L^q(\Omega)$. The fractional powers A^θ, $0 < \theta < 1$, are well-defined.

(b) γ denotes any continuous operator [Trie, Wahl]

$$\gamma : \mathcal{D}(A^\sigma) \equiv W^{2\sigma,q}(\Omega) \to U \equiv L^q(\Omega),\ 2\sigma = \frac{1}{q} + \varepsilon \quad (2.3)$$

in particular the trace operator

$$\gamma\psi \equiv \psi|_\Gamma \in L^q(\Gamma),\ \psi \in W^{2\sigma,q}(\Omega). \quad (2.4)$$

Thus, the (feedback) operator F defined in (2.1d) satisfies

$$F : \mathcal{D}(A^\sigma) \equiv W^{2\sigma,q}(\Omega) \to U \equiv L^q(\Omega),\ 2\sigma = \frac{1}{q} + \varepsilon \quad (2.5)$$

as well, for all vectors $w_k \in L^{q'}(\Gamma), g \in L^q(\Gamma), \dfrac{1}{q} + \dfrac{1}{q'} = 1$, where $(\ ,\ )_\Gamma$ denotes the duality paring between $L^q(\Gamma)$ and $L^{q'}(\Gamma)$.

(c) We introduce the Neumann (Green) map [LT.4]

$$Gg \equiv \varphi \iff \left\{ (\Delta - I)\varphi \equiv 0 \text{ in } \Omega,\ \frac{\partial \varphi}{\partial \nu} = g \text{ on } \Gamma \right\} \quad (2.6a)$$

$$G : U \equiv L^q(\Omega) \to W^{1+1/q,q}(\Omega) \subset \mathcal{D}(A^{\alpha_0}),\ \alpha_0 = \frac{1}{2} + \frac{1}{2q} - \varepsilon. \quad (2.6b)$$

(d) We observe from (2.3) and (2.6b) that

$$\sigma = \frac{1}{2q} + \frac{\varepsilon}{2} < \alpha_0 = \frac{1}{2} + \frac{1}{2q} - \varepsilon \quad (2.7)$$

The abstract model. As is well known, we can rewrite (2.1a) as

$$y_t = (\Delta - I)y = (\Delta - 1)(y - Gf),\ \text{since } (\Delta - 1)(Gf) \equiv 0 \text{ in } \Omega \quad (2.8)$$

by (2.6a), recalling f in (2.1c). Moreover

$$\frac{\partial(y-Gf)}{\partial\nu} = \frac{\partial y}{\partial\nu} - \frac{\partial(Gf)}{\partial\nu} = f - f \equiv 0 \text{ on } \Gamma \tag{2.9}$$

and so $(y-Gf)$ satisfies the boundary conditions of the operator A in (2.2b). In conclusion, recalling $f = Fy$ from (2.1d) we can rewrite (2.8) as

$$y_t = -A(I-GF)y = A_F y \tag{2.10}$$

which is the abstract model on $Y \equiv L^q(\Omega)$ of the original PDE feedback model (2.1a-d). We now verify that the abstract model (2.10) for (2.1a-d) satifies all abstract assumptions of Section 1.

(**H.1**)is satisfied since $Y \equiv L^q(\Omega)$, $1 < q < \infty$ is reflexive Banach space. (**H.2**)is satisfied since the operator $-A$ in (2.2a) is the maximal dissipative generator of a C_0-contraction semigroup e^{-At} on Y, $t \geq 0$, which possesses the maximal $L^p(0,T;Y)$-regularity property, $0 < T < \infty$, $1 < p < \infty$. (**H.3**)is satisfied since $U = L^q(\Gamma)$ is a Banach space $A^{\alpha_0}G \in \mathcal{L}(U;Y)$ from (2.6b), $\alpha_0 < 1$. (**H.4**)is satisfied by (2.5).
In conclusion: Problem (2.1a-d) satisfies all assumptions of Theorem 1.2, for $0 < T < \infty$, and hence $A_F \in MReg(L^p(0,T;L^q(\Omega)))$, $1 < p < \infty$, $1 < p < \infty$, $T < \infty$ with $A_F = -A(I-GF)$ in (2.10). This conclusion is true for all $w_k \in L^{q'}(\Gamma)$, $g_k \in L^q(\Gamma)$. Below we shall consider the case $T = \infty$.

Example # 2: We return to [LT.6, LPT.2] and consider the linearized Navier-Stokes problem over a bounded domain Ω in $\mathbb{R}^d$, $d = 2, 3$, with boundary $\partial\Omega \equiv \Gamma$ (after translation by the equilibrium solution, see [LPT.2, Eq (1.28)]))

$$\begin{cases} w_t - \nu_o \Delta w + L_e(w) + \nabla\chi = 0 & \text{in } Q \quad (2.11a) \\ \text{div } w = 0 & \text{in } Q \quad (2.11b) \\ w \equiv v \equiv \sum_{k=0}^{K} \langle \gamma w, p_k \rangle_\Gamma \, g_k \equiv Fw & \text{on } \Sigma \quad (2.11c) \\ w(0,x) = w_0(x) & \text{on } \Omega \quad (2.11d) \end{cases}$$

whose abstract version is given by

$$\frac{dw}{dt} = \mathcal{A}_q w - \mathcal{A}_q D\left(\sum_{k=1}^{K} \langle \gamma w, p_k \rangle_\Gamma \, g_k\right) \tag{2.12a}$$

$$= \mathcal{A}_q w - \mathcal{A}_q DFw = \mathcal{A}_q(I - DF) \equiv \mathbb{A}_{F,q} w. \tag{2.12b}$$

see [LPT.2, Eq (4.3)] with $m \equiv 0$ and a modified boundary control v. We have

$$Y \equiv L^q_\sigma(\Omega),\ q \geq 2,\ \mathcal{A}_q = -(\nu_o A_q + A_{o,q}),\ \mathcal{D}(\mathcal{A}_q) = \mathcal{D}(A_q) \subset L^q_\sigma(\Omega) \quad \text{[LPT.2, Eq (2.16)]} \tag{2.13}$$

$$A_q z = -P_q \Delta z,\ \mathcal{D}(A_q) = W^{2,q}(\Omega) \cap W_0^{1,q}(\Omega) \cap L^q_\sigma(\Omega) \quad \text{[LPT.2, Eq (2.14)]} \tag{2.14}$$

$$L_e(z) = (y_e \,.\, \nabla)z + (z \,.\, \nabla)y_e \quad \text{[LPT.2, Eq (1.9)]} \tag{2.15}$$

$$A_{o,q} z = P_q L_e(z) = P_q[(y_e \,.\, \nabla)z + (z \,.\, \nabla)y_e], \tag{2.16a}$$

$$\mathcal{D}(A_{o,q}) = \mathcal{D}(A_q^{1/2}) = W_0^{1,q}(\Omega) \cap L^q_\sigma(\Omega) \subset L^q_\sigma(\Omega). \quad \text{[LPT.2, Eq (2.15)]} \tag{2.16b}$$

$$L^q(\Omega) = L^q_\sigma(\Omega) \oplus G^q(\Omega) \quad \text{(Helmholtz direct sum decomposition)} \tag{2.17}$$

$$L^q_\sigma(\Omega) = \{g \in L^q(\Omega) : \div g = 0;\ g \cdot \nu = 0 \text{ on } \partial\Omega\}, \quad \text{[Ga]} \tag{2.18}$$

the solenoidal space. It is verified in [LPT.6] that all the assumptions of Theorem 1.2 are satisfied for the feedback operator $\mathbb{A}_{F,q} = \mathcal{A}_q(I - DF)$ in (2.12b) on $Y \equiv L^q_\sigma(\Omega)$. In particular, (**H.2**) is verified since the operator $\mathcal{A}_q$ in (2.13) has maximal L^p-regularity on $Y \equiv L^q_\sigma(\Omega)$, for $0 < T < \infty$. (**H.3**) is satisfied with

$$U \equiv U_q \equiv \{g \in L^q(\Gamma) :\ g \cdot \nu = 0 \text{ on } \Gamma\} \tag{2.19}$$

$$D : U_q \to W^{1/q,q}(\Omega) \cap L^q_\sigma(\Omega) \subset \mathcal{D}(A_q^{1/2q-\varepsilon}) \tag{2.20a}$$

$$\text{or } A_q^{1/2q-\varepsilon} D \in \mathcal{L}(U_q, L^q_\sigma(\Omega)), \quad \sigma_0 = \frac{1}{2q} - \varepsilon \tag{2.20b}$$

(**H.4**) is satisfied by taking

$$\gamma : \text{ continuous } \mathcal{D}(A_q^\sigma) \subset Y \equiv L_\sigma^q(\Omega) \to U \text{ with } 0 < \sigma < \sigma_0 = \frac{1}{2q} - \varepsilon \tag{2.21}$$

so that then

$$F : \text{ continuous } \mathcal{D}(A_q^\sigma) \subset Y \to U \tag{2.22}$$

as well. Then we take $p_k \in L^{q'}(\Gamma)$ and $g \in L^q(\Gamma)$, $\frac{1}{q} + \frac{1}{q'} = 1$. Then all the assumptions of Theorem 1.2 are satisfied for the feedback operator $\mathbb{A}_{_{F,q}} = \mathcal{A}_q(I - DF)$ in (2.12b). See [Sol.1, Sol.2, Sol.3, Sol.4] for open-loop problems.

2.2. Case $T = \infty$

We return to Example # 1, except that, to make the problem more significant, we replace (2.1a) by the canonical equation

$$\frac{\partial y(t,x)}{\partial t} = (\Delta + k^2)y(t,x) \tag{2.23}$$

k^2 large, while keeping Eqts (2.1b-c). Thus, for $f \equiv 0$, the corresponding free dynamics operator

$$A\varphi = (\Delta + k^2)\varphi, \quad Y \equiv L^q(\Omega) \supset \mathcal{D}(A) = \left\{\varphi \in W^{2,q}(\Omega), \ \left.\frac{\partial \varphi}{\partial \nu}\right|_\Gamma = 0\right\} \tag{2.24}$$

is the generator of a s.c. analytic semigroup on Y which is unstable and possesses maximal $L^p(0,T;Y)$-regularity, $T < \infty$. We take the boundary vectors $g_k \in L^q(\Gamma)$ to be linearly independent. According to Theorem 1.1(b), or the basis of the analysis of Example # 1 (k^2 rather than -1 is irrelevant), we only need to verify the additional assumption that, for suitable vectors $w_k \in L^{q'}(\Gamma)$, $g_k \in L^q(\Gamma)$, the semigroup $T_F(t) = e^{A_F t}$, $A_F = -A(I - GF)$ in (2.10), is exponentially stable

$$\left\|e^{A_F t}\right\|_{\mathcal{L}(Y)} \equiv \left\|T_F(t)\right\|_{\mathcal{L}(Y)} \leq Me^{-\delta t}, \quad t \geq 0, \ \delta > 0, \ Y \equiv L^q(\Omega). \tag{2.25}$$

This statement amounts to saying that the original boundary homogeneous problem (2.23), (2.1a-d) which with $f \equiv 0$ is unstable (i.e. it has finitely many unstable eigenvalues on $\mathcal{C}^+ = \{\lambda \in \mathcal{C} : \text{Re}\,\lambda \geq 0\}$) can be uniformly stabilized by a finite dimensional feedback control $f(t,\xi) =$ RHS of (2.1c), with suitable boundary vectors $g_k \in L^q(\Gamma)$ and boundary sensors $w_k \in L^{q'}(\Gamma)$. This problem was originally studied in early 1980s, see [LT.2, LT.3, Tr.1, Tr.2, Tr.3] and references therein. The vectors w_k have to be chosen to satisfy the algebraic condition

$$\begin{aligned}\text{rank } W_k = \ell_k = \ &\text{algebraic/geometric multiplicity of the unstable eigenvalue } \lambda_k \\ &\text{of the self-adjoint operator } A \text{ in (2.24)}\end{aligned} \tag{2.26}$$

where

$$W_k = \begin{bmatrix} \langle w_1, \Phi_{k1}\rangle_\Gamma, \langle w_1, \Phi_{k2}\rangle_\Gamma & \dots & \langle w_1, \Phi_{k\ell_k}\rangle_\Gamma \\ \langle w_2, \Phi_{k1}\rangle_\Gamma, \langle w_2, \Phi_{k2}\rangle_\Gamma & \dots & \langle w_2, \Phi_{k\ell_k}\rangle_\Gamma \\ \vdots & & \vdots \\ \langle w_K, \Phi_{k1}\rangle_\Gamma, \langle w_K, \Phi_{k2}\rangle_\Gamma & \dots & \langle w_K, \Phi_{k\ell_k}\rangle_\Gamma \end{bmatrix} \tag{2.27}$$

$\langle \cdot, \cdot \rangle_\Gamma$ duality pair, where $\{\Phi_{k1}, \dots, \Phi_{k\ell_k}\}$ are the normalized eigenvectors in Y of the unstable eigenvalues λ_k of the operator A in (2.24). Condition (2.27) can always be satisfied by infinite choices of the vectors $w_1, \dots, w_K$, since for every λ_k, the Dirichlet traces $\{\Phi_{k1}|_\Gamma, \Phi_{k2}|_\Gamma, \dots, \Phi_{k\ell_k}|_\Gamma\}$ are linearly independent [Tr.4, Tr.5, Tr.6].

It is known [LT.3] that if Ω is either a d-sphere or a d-parallelepiped, is it always possible to select boundary vectors g_k, $k = 1, \dots, k$ such that the exponential decay (2.25) holds true [LT.3] and hence Theorem 1.2 holds true for $T = \infty$ for these special geometries. For other geometries, $d \geq 2$, technical conditions are available which cannot be recalled here for brevity of exposition. We refer to the reference [LT.5] Moreover, if $d = 1$, the uniform stability (2.25) is impossible if A has at least 3 unstable eigenvalues, with only to boundary vectors g_1, g_2 at $x = 0$, or $x = 1$ with $\Omega = (0,1)$ [LT.3]. See [LT.1] for Dirichlet boundary feedback problems.

Acknowledgements

The research of Irena Lasiecka and Roberto Triggiani was partially supported by the National Science Foundation under Grant #2205508. Furthermore, the research of Irena Lasiecka was partially supported by the NCS, grant Opus, Agreement UMO-2023/49/B/ST1/04261. The research of Buddhika Priyasad was partially supported by the YSF offered by University of Konstanz under the project number: FP 638/23.

References

[CV] P. Cannarsa, V. Vespri, On Maximal L^p regularity for the abstract Cauchy problem, Boll. Un. Mat. Ital B (6) 5 (1986) n 1, 165-175.

[DaG] G. DaPrato, P. Grisvard, Maximal regularity for evolution equations by interpolation and extrapolation, Journal of Functional Analysis, Volume 58, Issue 2, 1984, 107-124.

[DaV] G. DaPrato, V. Vespri, Maximal L^p regularity for elliptic equations with unbounded coefficients, NonLinear Analysis 49 (2002) n 6 Ser A: Theory Methods, 747-755.

[DHP] R. Denk, M. Hieber, J. Prüss, Optimal $L^p - L^q$-regularity for parabolic problems with inhomogeneous boundary data, Math. Z., 257 (2007), pp. 193-224.

[Dore.1] G. Dore, L^p regularity for abstract differential equations, *Functional Analysis and Related Topics*, 1991 (Kyoto), Springer-Berlin, pp 25-38.

[Dore.2] G. Dore, Maximal regularity in L^p spaces for an abstract Cauchy problem, *Advances in Differential Equations*, 2000.

[Fur] A. Fursikov, Stabilization for the 3D Navier–Stokes system by feedback boundary control, *DCDS* 10 (2004), 289–314.

[Ga] G. P. Galdi, An Introduction to the Mathematical Theory of the Navier-Stokes Equations. *Springer-Verlag New York*, 2011.

[HNVW.1] T. Hytönen, J. van Neerven, M. Veraar, L. Weis, Analysis in Banach Spaces, Volume 1 & Volume 2, Springer, 2016.

[HNVW.2] T. Hytönen, J. van Neerven, M. Veraar, L. Weis, Analysis in Banach Spaces, Volume 3, Springer, 2023.

[KW.1] P. C. Kunstmann, L. Weis, Perturbation theorems for maximal L^p-regularity *Annali della Scuola Normale Superiore di Pisa - Classe di Scienze,* Série 4 : Volume 30 (2001) no. 2 , p. 415-435

[KW.2] P. C. Kunstmann, L. Weis, Maximal L^p-regularity for Parabolic Equations, Fourier Multiplier Theorems and H^∞-functional Calculus *Functional Analytic Methods for Evolution Equations, Lecture Notes in Mathematics, vol 1855. Springer, Berlin, Heidelberg* pp 65-311

[Las] I. Lasiecka, Unified theory for abstract parabolic boundary problems—a semigroup approach, *Appl Math Optim* 6, 287–333 (1980). https://doi.org/10.1007/BF01442900.

[LPT.1] I. Lasiecka, B. Priyasad, R. Triggiani, Uniform Stabilization of Navier–Stokes Equations in Critical L^q-Based Sobolev and Besov Spaces by Finite Dimensional Interior Localized Feedback Controls. *Appl. Math Optim.* (2019). https://doi.org/10.1007/s00245-019-09607-9

[LPT.2] I. Lasiecka, B. Priyasad, R. Triggiani, Uniform stabilization of 3D Navier-Stokes equations in critical Besov spaces with finite dimensional, tangential-like boundary, localized feedback controllers, Arch Rational Mech Anal 241, 1575–1654 (2021). https://doi.org/10.1007/s00205-021-01677-w.

[LPT.3] I. Lasiecka, B. Priyasad, R. Triggiani, Uniform stabilization of Boussinesq systems in critical L^q-based Sobolev and Besov spaces by finite dimensional interior localized feedback controls, *Discrete & Continuous Dynamical Systems - B*, 25, 10, 4071, 4117, 2020-6-15, 1531-3492_2020_10_4071.

[LPT.4] I. Lasiecka, B. Priyasad, R. Triggiani, Finite dimensional boundary uniform stabilization of the Boussinesq system in Besov spaces by critical user of Carleman estimate-based inverse theory, *Journal of Inverse and Ill-posed Problems*. vol. 30, no. 1, 2022, pp. 35-79. https://doi.org/10.1515/jiip-2020-0132.

[LPT.5] I. Lasiecka, B. Priyasad, R. Triggiani, Maximal L^p-regularity for an abstract evolution equation with applications to closed-loop boundary feedback control problems, *Journal of Differential Equations* Volume 294, 2021, Pages 60-87, ISSN 0022-0396, https://doi.org/10.1016/j.jde.2021.05.046.

[LPT.6] I. Lasiecka, B. Priyasad, R. Triggiani, Maximal L^p-regularity of an abstract evolution equation: application to closed-loop feedback problems, with boundary controls and boundary sensors, submitted.

[LPT.7] I. Lasiecka, B. Priyasad, R. Triggiani, Uniform Stabilization in Besov spaces with arbitrary decay rates of the Magnetohydrodynamic System by finite-dimensional interior localized feedback controllers, in preparation.

[LT.1] I. Lasiecka, R. Triggiani, Stabilization and structural assignment of Dirichlet boundary feedback parabolic equations, *SIAM J. Control Optimiz.*, 21 (1983), 766-803.

[LT.2] I. Lasiecka, R. Triggiani, Feedback semigroups and cosine operators for boundary feedback parabolic and hyperbolic equations, *J. Diff. Eqns.*, 47 (1983), 246-272.

[LT.3] I. Lasiecka, R. Triggiani, Stabilization of Neumann boundary feedback parabolic equations: The case of trace in the feedback loop, *Appl. Math. Optimiz.*, 10 (1983), 307-350.

[LT.4] I. Lasiecka, R. Triggiani, Control Theory for Partial Differential Equations: Continuous and Approximation Theories, Vol. 1, Abstract Parabolic Systems (680 pp.), *Encyclopedia of Mathematics and its Applications Series*, Cambridge University Press, January 2000.

[LT.5] I. Lasiecka, R. Triggiani, Uniform Stabilization with Arbitrary Decay Rates of the Oseen Equation by Finite-Dimensional Tangential Localized Interior and Boundary Controls. *Semigroups of Operators -Theory and Applications*, Proms 113, 2015, 125-154.

[LT.6] I. Lasiecka, R. Triggiani, Stabilization to an Equilibrium of the Navier-Stokes Equations with Tangential Action of Feedback Controllers. *Nonlinear Analysis*, 121 (2015), 424-446.

[Pazy] A. Pazy, Semigroups of Linear Operators and Applications to Partial Differential Equations, *Springer-Verlag*, 1983.

[Sim] L. De Simon, Un'applicazione della teoria degli integrali singolari allo studio delle equazioni differenziali lineari astratte del primo ordine, *Rendiconti del Seminario Matematico della Universitá di Padova* (1964), Volume: 34, page 205-223.

[Sol.1] V. A. Solonnikov, *Estimates of the solutions of a nonstationary linearized system of Navier-Stokes equations*, A.M.S. Translations, 75 (1968), 1-116.

[Sol.2] V. A. Solonnikov, Estimates for solutions of non-stationary Navier-Stokes equations, *J. Sov. Math.*, 8, 1977, pp 467-529.

[Sol.3] V. A. Solonnikov, On the solvability of boundary and initial-boundary value problems for the Navier-Stokes system in domains with noncompact boundaries. *Pacific J. Math.* 93 (1981), no. 2, 443-458. https://projecteuclid.org/euclid.pjm/1102736272.

[Sol.4] V. A. Solonnikov, L^p-Estimates for Solutions to the Initial Boundary-Value Problem for the Generalized Stokes System in a Bounded Domain, *J. Math. Sci.*, Volume 105, Issue 5, pp 2448–2484.

[Tr.1] R. Triggiani, On the Stabilizability Problem of Banach Spaces, *J. Math. Anal. Appl. 52 303-403*, 1975.

[Tr.2] R. Triggiani, Well-posedness and regularity of boundary feedback systems, *J. Diff. Eqns.*, 36(1980), 347-362.

[Tr.3] R. Triggiani, Boundary feedback stabilizability of parabolic equations, *Appl. Math. Optimiz.* 6 (1980), 201–220.

[Tr.4] R. Triggiani, Linear independence of boundary traces of eigenfunctions of elliptic and Stokes Operators and applications, invited paper for special issue, *Applicationes Mathematicae* 35(4) (2008), 481–512, Institute of Mathematics, Polish Academy of Sciences.

[Tr.5] R. Triggiani, Unique continuation of boundary over-determined Stokes and Oseen eigenproblems, *Discrete & Continuous Dynamical Systems - S*, Vol. 2 , N. 3, Sept 2009, 645-677.

[Tr.6] R. Triggiani, Unique Continuation from an Arbitrary Interior Subdomain of the Variable-Coefficient Oseen Equation. *Nonlinear Analysis Theory, Meth. & Appl.*, (17)2009, 4967-4976.

[Trie] H. Triebel, Interpolation Theory, Function Spaces, Differential Operators. *Bull. Amer. Math. Soc. (N.S.) 2, no. 2, 339-345* , 1980.

[Ves] V. Vespri, Regolaritá massimale in L^p per il problema di Cauchy astratto e regolaritá $L^p(L^q)$ per operatori parabolici, in: L. Modica (Ed.) "*Atti del convegno su equazioni differenziali e calcolo delle variazioni*"; Pisa, 1985, 205-213.

[Wahl] W. von Wahl, The Equations of Navier-Stokes and Abstract Parabolic Equations. *Springer Fachmedien Wiesbaden, Vieweg+Teubner Verlag*, 1985.

[Weis] L. Weis, A new approach to maximal Lp-regularity. *In Evolution Equ. and Appl. Physical Life Sci.,* volume 215 of Lect. Notes Pure and Applied Math., pages 195–214, New York, 2001. Marcel Dekker.

FRENCH-GERMAN-SPANISH CONFERENCE ON OPTIMIZATION
Gijón, June 18-21, 2024
(pp. 102-107)

On some stochastic aspects of stochastic elliptic inverse problems

Akhtar A. Khan[1], **Miguel Sama**[2], **Hans-Jörg Starkloff**[3]

1. aaksma@rit.edu *Rochester Institute of Technology, Rochester, USA*
2. msama@ind.uned.es *Universidad Nacional de Educación a Distancia Madrid, Spain*
3. hjstark@math.tu-freiberg.de *Technische Universität Bergakademie Freiberg, Germany*

Abstract

In the article conditions are investigated which allow a direct transfer of results for deterministic elliptic variational problems to corresponding results for stochastic problems. Hereby measurability issues play a certain role and are discussed to some extend.

1. Introduction

Elliptic variational equations play an important role in mathematics, for theoretical investigations as well as for applications. This is related to direct problems and also to inverse problems. One aim of the present contribution consists in a discussion about possibilities to use the existing powerful abstract theory for deterministic problems also for stochastic problems, i.e., when some deterministic parameters in the problem under consideration are substituted by random ones.

As a prototypical example one can mention the deterministic elliptic boundary value problem on a well-behaved bounded domain $D \subset \mathbb{R}^n$ given by

$$-\nabla \cdot (a(x)\nabla u(x)) = f(x) \text{ in } D, \quad u(x) = 0, \text{ on } \partial D \tag{1.1}$$

with suitable functions a, f, such that with real constants c_1, c_2 it holds $0 < c_1 \leq a(x) \leq c_2 < \infty$ for all $x \in D$. Hereby the direct problem consists in determining u given a and f, whereas a typical inverse problem consists in finding a for given f and u. For a mathematical treatment an abstract formulation as a variational equation is advantageous, i.e., for all $v \in \mathrm{H}_0^1(D)$ it holds

$$\int_D a(x)\nabla u(x)\nabla v(x)\,\mathrm{d}x = \int_D f(x)v(x)\,\mathrm{d}x. \tag{1.2}$$

The left hand side of (1.2) determines an elliptic trilinear form and so results from functional analysis can be used to get results for direct and inverse problems related to this example.

Incorporating an existing uncertainty for the real system under investigation one can use a corresponding stochastic model. So for examples the functions a and f can be assumed to be random functions or random variables in suitable functional spaces. Then the question arises under which conditions one can use directly the results from the abstract theory in order to get corresponding results for the stochastic model.

1.1. Notation

In the article all vector spaces are non-trivial real vector spaces. For a Banach space $\mathbf{X}$ the topological dual space is denoted by $\mathbf{X}^*$, the norm by $\|\cdot\|_{\mathbf{X}}$ and the duality pairing with $\langle x^*|x\rangle_{\mathbf{X}}$ with $x^* \in \mathbf{X}^*, x \in \mathbf{X}$. If $\mathbf{X}$ is a Hilbert space the scalar product is denoted by $\langle\cdot,\cdot\rangle_{\mathbf{X}}$.

All random variables are defined on one underlying probability space $(\Omega, F, \mathbb{P})$ with expectation operator $\mathbb{E}[\cdot]$. $\mathbf{1}_B$ denotes the indicator function of a set B, i.e., $\mathbf{1}_B(x) = 1$ if $x \in B$ and 0 otherwise.

2. Deterministic elliptic variational equations

The following notations and assumptions are used throughout the article.

Assumption 2.1 *Let $\mathbf{V}$ be a Hilbert space, $\mathbf{B}$ be a Banach space and $\emptyset \neq K \subset \mathbf{B}$ be a convex closed set with non-empty interior. Also let $T : \mathbf{B} \times \mathbf{V} \times \mathbf{V} \to \mathbb{R}$ be a trilinear map (i.e., a map, linear in each of the three arguments with the other two being fixed) which is symmetric, i.e., $T(a,u,v) = T(a,v,u)$ for all $a \in \mathbf{B}, u, v \in \mathbf{V}$; continuous, i.e., there exists $\beta > 0$ such that $|T(a,u,v)| \leq \beta\|a\|_{\mathbf{B}}\|u\|_{\mathbf{V}}\|v\|_{\mathbf{V}}$ for all $a \in \mathbf{B}, u, v \in \mathbf{V}$; and elliptic on the set K, i.e., there exists $\alpha > 0$ such that $T(a,u,u) \geq \alpha\|u\|_{\mathbf{V}}^2$ for all $a \in K$ and $u \in \mathbf{V}$.*

The famous Lax-Milgram lemma states the unique solvability of the abstract variational equation if Assumption 2.1 holds (see e.g. [1], Theorem 1.1.3, Remark 1.1.3).

Theorem 2.2 *For given $a \in K, f \in \mathbf{V}^*$ under Assumption 2.1 the abstract elliptic variational equation*

$$\forall\, v \in \mathbf{V} : T(a,u,v) = \langle f|v\rangle_{\mathbf{V}} \tag{2.1}$$

has a unique solution $u = u(a,f) \in \mathbf{V}$ and it holds $\|u(a,f)\|_{\mathbf{V}} \leq \alpha^{-1}\|f\|_{\mathbf{V}^*}$.

Also the Lemma of Céa (see e.g. [1], Theorem 2.4.1) will be important in the following.

Theorem 2.3 *Let Assumption 2.1 be fulfilled, $a \in K, f \in \mathbf{V}^*$, let $\mathbf{V}_1$ be a closed subspace of $\mathbf{V}$ and consider the elliptic variational equation* (2.1) *on $\mathbf{V}$ with unique solution $u(a,f) \in \mathbf{V}$ and on $\mathbf{V}_1$ with unique solution $u_1(a,f) \in \mathbf{V}_1$. Then it holds*

$$\|u(a,f) - u_1(a,f)\|_{\mathbf{V}} \leq \beta\alpha^{-1} \inf_{v_1 \in \mathbf{V}_1} \|u(a,f) - v_1\|_{\mathbf{V}}. \tag{2.2}$$

Corollary 2.4 *Let Assumption 2.1 be fulfilled with a separable Hilbert space $\mathbf{V}$, $a \in K$, $f \in \mathbf{V}^*$ and let $(\mathbf{V}_n)_{n\in\mathbb{N}}$ be an increasing sequence of closed subspaces of $\mathbf{V}$ with dense union in $\mathbf{V}$. Then it holds for the sequence $(u_n(a,f))_{n\in\mathbb{N}}$ of solutions to the elliptic variational equations* (2.1) *on $\mathbf{V}_n$ correspondingly $\lim_{n\to\infty} u_n(a,f) = u(a,f)$.*

For a mathematical treatment of inverse problems for deterministic elliptic variational equations like (2.1) an investigation of the parameter to solution map $K \ni a \mapsto u = u(a,f) \in \mathbf{V}$ is useful. The following two results can be found for example in [4].

Theorem 2.5 *Let Assumption 2.1 be fulfilled, $f \in \mathbf{V}^*$ and consider the abstract elliptic variational equation* (2.1)*. Then the parameter to solution map $K \ni a \mapsto u = u(a,f) \in \mathbf{V}$ is Lipschitz continuous and it holds for $a_1, a_2 \in K$*

$$\|u(a_1,f) - u(a_2,f)\|_{\mathbf{V}} \leq \beta\alpha^{-2}\|f\|_{\mathbf{V}^*}\|a_1 - a_2\|_{\mathbf{B}}. \tag{2.3}$$

Theorem 2.6 *Let Assumption 2.1 be fulfilled, $f \in \mathbf{V}^*$ and consider the abstract elliptic variational equation* (2.1)*. Then for each a in the interior of K the parameter to solution map $K \ni a \mapsto u = u(a,f) =: F(a) \in \mathbf{V}$ is differentiable at a, and $\delta u := DF(a)\delta a \in \mathbf{V}$ with $\delta a \in \mathbf{B}$ is the unique solution to the variational equation*

$$\forall\, v \in \mathbf{V} : T(a,\delta u,v) = -T(\delta a, u(a,f), v). \tag{2.4}$$

Moreover,

$$\|DF(a)\| \leq \beta\alpha^{-1}\|F(a)\|_{\mathbf{V}} \leq \beta\alpha^{-2}\|f\|_{\mathbf{V}^*}. \tag{2.5}$$

Also higher order derivatives exist and they are defined by corresponding elliptic variational equations (see e.g. [4]).

If some data $z \in \mathbf{V}$ is given approximating the solution $u(a,f)$ for a given element $f \in \mathbf{V}^*$ and an unknown element $a \in K$, the inverse problem of identifying $a \in K$ can be tackled for example minimizing some error functional. As was shown for example in [4], using for this purpose the energy least-squares functional

$$J(a;z,f) := \frac{1}{2}T(a,u(a,f)-z,u(a,f)-z) \tag{2.6}$$

has the advantage that it is smooth and convex.

Theorem 2.7 *Let Assumption 2.1 be fulfilled, $f \in \mathbf{V}^*$ and consider the abstract elliptic variational equation* (2.1)*. For given data $z \in \mathbf{V}$, energy least-squares functional $K \ni a \mapsto J(a;f) \geq 0$ is smooth and convex on K.*

Due to the inherent ill-posedness usually one minimizes the regularized energy least-squares functional

$$J_\kappa(a;z,f) := \frac{1}{2}T(a,u(a,f)-z,u(a,f)-z) + \frac{\kappa}{2}\|a\|_{\mathbf{B}}^2 \tag{2.7}$$

with $\kappa > 0$. Then under additional assumptions one can show, that there is a unique solution for the minimization problem for the regularized energy least-squares functional (see e.g. Sections 3 and 4 of [4]).

3. Pathwise solutions to stochastic elliptic variational equations and measurability issues

In situations with a relevant uncertainty one has to consider a multitude of possible elements a and f (hence also solutions $u(a,f)$), which usually one describes using parametrized families of such elements. The non-empty parameter set will be denoted by Ω here, its elements by ω. From Theorem 2.2 one easily concludes

Proposition 3.1 *Let Assumption 2.1 be fulfilled, let $\Omega \neq \emptyset$, and assume for the parametrized mappings $a : \Omega \ni \omega \mapsto a(\omega) \in K \subset \mathbf{B}$ and $f : \Omega \ni \omega \mapsto f(\omega) \in \mathbf{V}^*$. Then the parametrized elliptic variational equation*

$$\forall\, v \in V : T(a(\omega), u, v) = \langle f(\omega) | v \rangle_{\mathbf{V}} \tag{3.1}$$

has for all $\omega \in \Omega$ a unique solution $u(\omega) := u(a(\omega), f(\omega)) \in \mathbf{V}$ and it holds $\|u(\omega)\|_{\mathbf{V}} \leq \alpha^{-1} \|f(\omega)\|_{\mathbf{V}^*}$.

Now assume that one wants to use a stochastic description for an uncertainty quantifaction. Then the parameter set Ω is the basic set of a probability space $(\Omega, F, \mathbb{P})$ with a σ-algebra F of subsets of Ω (the measurable sets) and a probability measure $\mathbb{P}$ defined on F. In order to use results from probability theory one has to ensure that relevant mappings are measurable. Here mappings from the underlying probability space into a Banach space are of interest. As a first step we define two basic σ-algebras in a general Banach space X. The following definitions and propositions can be found for example in [2] and [6].

Definition 3.2 Let $\mathbf{X}$ be a Banach space. The σ-algebra generated by the set of all open subsets of $\mathbf{X}$ is the Borel-σ-algebra $B(\mathbf{X})$, whereas the σ-algebra generated by the set of all continuous linear functionals on $\mathbf{X}$ is the cylindrical σ-algebra, which will be denoted by $C(\mathbf{X})$.

Proposition 3.3 *Let $\mathbf{X}$ be a Banach space. Then it holds $C(\mathbf{X}) \subseteq B(\mathbf{X})$. If $\mathbf{X}$ is a separable Banach space it holds $C(\mathbf{X}) = B(\mathbf{X})$.*

In general there are different types of definitions of measurability. One distinguishes descriptive and constructive definitions of measurability. Descriptive definitions of measurability are characterized by the property, that inverse images of measurable set in the image space are measurable sets in the domain of the mapping. They are taylored for the transfer of measures from the domain into the image space and allow for example the definition of the distribution of random variables. Constructive definitions of measurability are characterized by the property, that corresponding measurable maps are limits in an appropriate sense of a sequence of simple measurable mappings. They allow for example an efficient definition of integrals (expectations).

For mappings from a probability space $(\Omega, F, \mathbb{P})$ into a Banach space $\mathbf{X}$ the most important measurability concepts are the following three. Hereby the first two are descriptive definitions, the last is a constructive one.

Definition 3.4 Let $(\Omega, F, \mathbb{P})$ be a probability space, $\mathbf{X}$ be a Banach space and $\Omega \ni \omega \mapsto X(\omega) \in \mathbf{X}$ be a mapping.

1. X is Borel measurable, if it is $F - B(\mathbf{X})$-measurable, i.e., $X^{-1}(B) \in F$ for all $B \in B(\mathbf{X})$.
2. X is weakly measurable, if it is $F - C(\mathbf{X})$-measurable, i.e., for all $x^* \in \mathbf{X}^*$ the mapping $\Omega \ni \omega \mapsto \langle x^* | X(\omega) \rangle_{\mathbf{X}} \in \mathbb{R}$ is Borel measurable.
3. X is strongly measurable, if it is the pointwise limit of a sequence of finitely-valued Borel measurable mappings, i.e.,

$$\forall\, \omega \in \Omega : X(\omega) = \lim_{n \to \infty} X_n(\omega) \quad \text{with} \quad X_n = \sum_{k=1}^{N_n} x_{n,k} \mathbf{1}_{B_{n,k}} \tag{3.2}$$

and $N_n \in \mathbb{N}$, $x_{n,k} \in \mathbf{X}$, $B_{n,k} \in B(\mathbf{X})$, $k, n \in \mathbb{N}$.

One can remark that traditionally in the definition of strong measurability the $\mathbb{P}$-almost sure convergence is used instead of the pointwise convergence. Then the corresponding maps are not necessarily Borel measurable, which may cause some (light) problems. That's why the definition from [2] is used here.

The following relations between the measurability concepts of Banach space-valued mappings are often used.

Proposition 3.5 *Assume there are given a Banach space $\mathbf{X}$, a probability space $(\Omega, F, \mathbb{P})$ and a mapping $X : \Omega \to \mathbf{X}$. Then it holds:*

(i) *X is strongly measurable* $\Rightarrow$ *X is Borel measurable;*

(ii) *X is Borel measurable* $\Rightarrow$ *X is weakly measurable.*

For a separable Banach space $\mathbf{X}$ *it holds additionally:*

(iii) *X is Borel measurable* $\Rightarrow$ *X is strongly measurable;*

(iv) *X is weakly measurable* $\Rightarrow$ *X is Borel measurable.*

Hence in a separable Banach space all three basic measurability concepts from above coincide. This is in general not true for non-separable Banach spaces.

In order to apply the abstract theory for stochastic problems spaces of random variables should be vector spaces. In general this cannot be achieved with all the measurability concepts which were introduced earlier.

Proposition 3.6 *Assume there are given a Banach space* $\mathbf{X}$ *and a probability space* $(\Omega, F, \mathbb{P})$*. Then the set of all weakly measurable mappings from* Ω *to* $\mathbf{X}$ *and the set of all strongly measurable mappings from* Ω *to* $\mathbf{X}$ *are vector spaces. If* $\mathbf{X}$ *is a separable Banach space, this holds also for the set of all Borel measurable mappings from* Ω *to* $\mathbf{X}$*.*

A further necessary condition for using Banach spaces of (equivalence classes) of random variables is the measurability of the norm of a random variable in a Banach space.

Proposition 3.7 *Assume there are given a Banach space* $\mathbf{X}$*, a probability space* $(\Omega, F, \mathbb{P})$ *and a mapping* $X : \Omega \to \mathbf{X}$*, which is Borel measurable or strongly measurable. Then the mapping* $\Omega \ni \omega \mapsto \|X(\omega)\|_{\mathbf{X}} \in [0, \infty)$ *is* $F - B([0, \infty))$*-measurable, i.e., a usual real-valued random variable. If* $\mathbf{X}$ *is a separable Banach space, this holds also if X is weakly measurable.*

From the results above it follows that the class of strongly measurable mappings ist the most appropriate class of mappings for using Banach spaces. In order to get really Banach spaces on should additionally identify mappings which are equal to each other $\mathbb{P}$-almost surely.

Definition 3.8 Assume there are given a Banach space $\mathbf{X}$, a probability space $(\Omega, F, \mathbb{P})$ and $p \in [1, \infty)$.

1. The set of all equivalence classes of strongly measurable mappings X from Ω to $\mathbf{X}$ with $\mathbb{E}\left[\|X\|_{\mathbf{X}}^p\right] < \infty$ is the Bochner space $\mathrm{L}^p(\Omega; \mathbf{X})$.
2. The Bochner space $\mathrm{L}^\infty(\Omega; \mathbf{X})$ is the set of all equivalence classes of strongly measurable mappings X from Ω to $\mathbf{X}$ with $\mathbb{P}(\|X\|_{\mathbf{X}} \leq M) = 1$ for some $M \geq 0$.

Proposition 3.9 *Under the assumptions of Definition 3.8 it holds*

(i) *The Bochner spaces* $\mathrm{L}^p(\Omega; \mathbf{X})$ *with* $p \in [1, \infty]$ *are Banach spaces with norm* $\|X\|_{\mathrm{L}^p(\Omega;\mathbf{X})} := \left(\mathbb{E}\|X\|_{\mathbf{X}}^p\right)^{1/p}$ *for* $1 \leq p < \infty$ *and* $\|X\|_{\mathrm{L}^\infty(\mathbf{X})} := \operatorname{esssup}_{\omega\in\Omega}\|X(\omega)\|_{\mathbf{X}}$ *for* $p = \infty$*, respectively.*

(ii) *If* $\mathbf{X}$ *is a Hilbert space the Bochner space* $\mathrm{L}^2(\Omega; \mathbf{X})$ *is a Hilbert space with scalar product* $\langle X_1, X_2\rangle_{\mathrm{L}^2(\Omega;\mathbf{X})} := \mathbb{E}\left[\langle X_1, X_2\rangle_{\mathbf{X}}\right]$.

After having collected some results related to measurability issues for Banach space-valued mappings defined on a probability space we consider the question of measurability of the parametrized solution of the parametrized elliptic variational equation from Proposition 3.1. As usual in probability theory mappings depending on the elementary elements ω are written also without indicating the arguments, for example writing $f(\cdot)$, etc.

Theorem 3.10 *Consider a parametrized elliptic variational equation as in Proposition 3.1 with a probability space* $(\Omega, F, \mathbb{P})$ *and a separable Hilbert space* $\mathbf{V}$*, i.e.,*

$$\forall\, v \in \mathbf{V} \quad T(a(\omega), u(\omega), v) = \langle f(\omega)|v\rangle_{\mathbf{V}}\,, \quad \mathbb{P} - a.s. \tag{3.3}$$

(i) *If* $f(\cdot)$ *is Borel measurable and if* $a(\cdot)$ *is weakly measurable, then the parametrized solution* $\Omega \ni \omega \mapsto u(\omega) = u(a(\omega), f(\omega)) \in \mathbf{V}$ *is Borel measurable and it holds* $\|u(\omega)\|_{\mathbf{V}} \leq \alpha^{-1}\|f(\omega)\|_{\mathbf{V}^*}$*,* $\mathbb{P}$*-a.s.*

(ii) *If additionally for* $p \geq 1$ *it holds* $\mathbb{E}\left[\|f(\cdot)\|_{\mathbf{V}^*}^p\right] < \infty$*, then it holds also* $\mathbb{E}\left[\|u(\cdot)\|_{\mathbf{V}}^p\right] < \infty$*.*

Proof Choose an orthonormal basis $(e_k)_{k\in\mathbb{N}}$ in the separable Hilbert space $\mathbf{V}$ and denote for $n \in \mathbb{N}$ by $\mathbf{V}_n$ the finite-dimensional subspace of $\mathbf{V}$ spanned by the first n basis elements. The parametrized solution of (3.3) $u_n(\cdot) = \sum_{k=1}^n \xi_k^{(n)}(\cdot)e_k$ on $\mathbf{V}_n$ (instead of $\mathbf{V}$) can be found determining $(\xi_1^{(n)}(\cdot), \dots, \xi_n^{(n)}(\cdot))^T$ by the parametrized system of linear equations

$$\sum_{k=1}^{n} T(a(\cdot), e_k, e_\ell)\xi_k^{(n)}(\cdot) = \langle f(\cdot)|e_\ell\rangle_{\mathbf{V}}, \quad \ell = 1, \dots, n, \tag{3.4}$$

with a measurable and regular matrix of the system $(T(a(\cdot), e_k, e_\ell))_{k,\ell=1,\dots,n}$ (it is strictly positive definite) and a measurable right-hand side. Hence there exists a unique measurable random solution $(\xi_1^{(n)}(\cdot), \dots, \xi_n^{(n)}(\cdot))^T$ of this system and a Borel measurable solution $u_n(\cdot)$. By the corollary of Céa's Lemma (see Corollary 2.4) these random solutions converge pointwise (for all parameter values ω) to the solution of (3.3) on the Hilbert space $\mathbf{V}$, so that this solution is also Borel measurable (as a pointwise limit of Borel measurable mappings in a separable Hilbert space). □

The previous theorem gives natural sufficient conditions for a pathwise solution of random elliptic variational equations. This result can be used for example for the justification of Monte Carlo algorithms for the solution of corresponding problems. But the concept of a pathwise solution is to weak for many other purposes, for example further investigations of certain properties of corresponding solutions or for solution algorithms using stochastic Galerkin methods. That's why one is attempting to determine stochastic solutions with the help of variational equations in spaces of random variables. This can be done with the help of the Bochner spaces, which were introduced earlier and include a stronger measurability condition on the coefficient $a(\cdot)$ in comparison with Theorem 3.10.

4. Integral stochastic elliptic variational equations as abstract elliptic variational equations

An integral stochastic elliptic variational equation is a special form of an abstract elliptic variational equation, where the Banach space and the Hilbert space are suitable Bochner spaces of equivalence classes of random variables in the corresponding deterministic functional spaces. Here we will deal only with the most straightforward variant of such problem formulations. In the following spaces of equivalence classes of random variables and elements of these spaces are denoted by a tilde in order to distinguish better the deterministic case and the stochastic case.

Assumption 4.1 *Additionally to Assumption 2.1 assume that* $\mathbf{V}$ *is a separable Hilbert space and* $(\Omega, F, \mathbb{P})$ *is a probability space. We use the notations* $\tilde{\mathbf{B}} := \mathrm{L}^\infty(\Omega; \mathbf{B})$, $\tilde{K} := \{\tilde{a} \in \tilde{\mathbf{B}} : \tilde{a}(\cdot) \in K\ \mathbb{P}-a.s.\}$, $\tilde{\mathbf{V}} := \mathrm{L}^2(\Omega; \mathbf{V})$, $\tilde{T} : \tilde{\mathbf{B}} \times \tilde{\mathbf{V}} \times \tilde{\mathbf{V}} \to \mathbb{R}$ *with* $\tilde{T}(\tilde{a}, \tilde{u}, \tilde{v}) := \mathbb{E}\,[T(\tilde{a}, \tilde{u}, \tilde{v})]$.

One can remark, that when K is bounded in $\mathbf{B}$ all equivalence classes of strongly measurable mappings $\Omega \to K$ belong to $\tilde{\mathbf{B}} := \mathrm{L}^\infty(\Omega; \mathbf{B})$.

Proposition 4.2 *Under Assumption 4.1 it holds:* $\tilde{\mathbf{B}}$ *is a Banach space,* $\tilde{\mathbf{V}}$ *is a Hilbert space,* $\tilde{K}$ *is a non-empty convex closed subset of* $\tilde{\mathbf{B}}$ *with non-empty interior. Moreover the mapping* $\Omega \ni \omega \mapsto T(\tilde{a}(\omega), \tilde{u}(\omega), \tilde{v}(\omega)) \in \mathbb{R}$ *is a random variable with finite expectation for* $\tilde{a} \in \tilde{\mathbf{B}}$, $\tilde{u} \in \tilde{\mathbf{V}}$, $\tilde{v} \in \tilde{\mathbf{V}}$ *(so that* $\tilde{T}$ *is well-defined) and* $\tilde{T}$ *is a trilinear form, which is symmetric, continuous and elliptic on the set* $\tilde{K}$.

Proof The first three properties are contained in Proposition 3.9 or can be checked easily. The measurability of $T(\tilde{a}, \tilde{u}, \tilde{v})$ follows from the continuity of T and the fact, that all random variables $\tilde{a}, \tilde{u}, \tilde{v}$ are pointwise limits of sequences of finitely-valued measurable mappings and for such random variables $T(\tilde{a}, \tilde{u}, \tilde{v})$ is measurable. Furthermore, $\mathbb{P}$-a.s. $|T(\tilde{a}(\omega), \tilde{u}(\omega), \tilde{v}(\omega))| \le \beta\|\tilde{a}(\omega)\|_{\mathbf{B}}\|\tilde{u}(\omega)\|_{\mathbf{V}}\|\tilde{v}(\omega)\|_{\mathbf{V}} \le \beta M\|\tilde{u}(\omega)\|_{\mathbf{V}}\|\tilde{v}(\omega)\|_{\mathbf{V}}$ for a random variable $\tilde{a} \in \tilde{\mathbf{B}}$ with $\mathbb{P}(\|\tilde{a}\|_{\mathbf{B}} \le M) = 1$ and $M > 0$. Hence using the Cauchy-Schwarz inequality

$$\mathbb{E}\,[|T(\tilde{a}, \tilde{u}, \tilde{v})|] \le \beta M \mathbb{E}\,[\|\tilde{u}\|_{\mathbf{V}}\|\tilde{v}\|_{\mathbf{V}}] \le \beta M \sqrt{\mathbb{E}\left[\|\tilde{u}\|_{\mathbf{V}}^2\right]\mathbb{E}\left[\|\tilde{v}\|_{\mathbf{V}}^2\right]} < \infty. \tag{4.1}$$

Similarly one shows the continuity of $\tilde{T}$. The symmetry of $\tilde{T}$ follows by its definition. Assume now $\tilde{a} \in \tilde{K}$. Then due to the ellipticity of T $\mathbb{P}$-a.s. $T(\tilde{a}(\cdot), \tilde{u}(\cdot), \tilde{u}(\cdot)) \ge \alpha\|\tilde{u}(\cdot)\|_{\mathbf{V}}^2$, so that $\mathbb{E}\,[T(\tilde{a}, \tilde{u}, \tilde{u})] \ge \alpha\mathbb{E}\left[\|\tilde{u}\|_{\mathbf{V}}^2\right] = \alpha\|\tilde{u}\|_{\mathrm{L}^2(\Omega;\mathbf{V})}^2$. □

From this one concludes easily the validity of the following main theorem.

Theorem 4.3 *Let Assumption 4.1 be fulfilled and let* $\tilde{a} \in \tilde{K} \subset \tilde{\mathbf{B}}$, $\tilde{f} \in \tilde{\mathbf{V}}^*$.

(i) The integral stochastic elliptic variational equation

$$\textit{find} \quad \tilde{u} \in \tilde{\mathbf{V}} : \forall\, \tilde{v} \in \tilde{\mathbf{V}} \quad \tilde{T}(\tilde{a}, \tilde{u}, \tilde{v}) = \langle \tilde{f} | \tilde{v} \rangle_{\tilde{\mathbf{V}}} =: \mathbb{E}\left[\langle \tilde{f} | \tilde{v} \rangle_{\mathbf{V}}\right] \tag{4.2}$$

is an abstract elliptic variational equation of the type (2.1) *in corresponding Bochner spaces of equivalence classes of random variables.*

(ii) All results for (2.1) *can directly be applied to* (4.2) *with a suitable change of spaces, norms etc.*

(iii) In particular for $\tilde{a} \in \tilde{K}$ *and* $\tilde{f} \in \tilde{\mathbf{V}}^*$ *there exists a unique solution* $\tilde{u} \in \tilde{\mathbf{V}}$ *to* (4.2) *and it holds* $\|\tilde{u}\|_{\tilde{\mathbf{V}}} \leq \alpha^{-1} \|\tilde{f}\|_{\tilde{\mathbf{V}}^*}$.

(iv) If strongly measurable random variables $\tilde{a}$ *and* $\tilde{f}$ *belong to* $\tilde{K}$ *and* $\tilde{\mathbf{V}}^*$*, respectively, the pathwise solution from Theorem 3.10 coincides* $\mathbb{P}$*-almost surely with the integral stochastic elliptic variational equation* (4.2)*. Then also the unique solution* $\tilde{u} \in \tilde{\mathbf{V}}$ *to* (4.2) *coincides* $\mathbb{P}$*-a.s. with the pathwise solution from Theorem 3.10.*

In particular also the results regarding the Lipschitz continuity of the parameter to solution map $\tilde{K} \ni \tilde{a} \mapsto \tilde{u}(\tilde{a}, \tilde{f}) \in \tilde{\mathbf{V}}$ or its derivatives or the convexity of the corresponding least-squares energy functional follow from Propositions 2.5, 2.6 and 2.7. Other proofs of these results and examples can be found e.g. in [5] or [3].

References

[1] Philippe G Ciarlet. *The finite element method for elliptic problems*. North-Holland Publishing Company, 1978.

[2] Donald L Cohn. *Measure theory*. Springer, 2013.

[3] Marc Dambrine, Akhtar A Khan, Miguel Sama, and Hans-Jörg Starkloff. Stochastic elliptic inverse problems. solvability, convergence rates, discretization, and applications. *Journal of Convex Analysis*, 30(3):851–885, 2023.

[4] Mark S Gockenbach and Akhtar A Khan. An abstract framework for elliptic inverse problems: Part 1. an output least-squares approach. *Mathematics and mechanics of solids*, 12(3):259–276, 2007.

[5] Baasansuren Jadamba, Akhtar A Khan, Miguel Sama, Hans-Jörg Starkloff, and Christiane Tammer. A convex optimization framework for the inverse problem of identifying a random parameter in a stochastic partial differential equation. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):922–952, 2021.

[6] Nikolai N Vakhania, Vazha I Tarieladze, and Sergei A Chobanyan. *Probability distributions on Banach spaces*. Springer, 2012.

French-German-Spanish Conference on Optimization
Gijón, June 18-21, 2024
(pp. 108-115)

A numerical solution approach for non-smooth optimal control problems based on the Pontryagin maximum principle

Daniel Wachsmuth
daniel.wachsmuth@uni-wuerzburg.de University of Würzburg

Abstract

We consider nonsmooth optimal control problems subject to a linear elliptic partial differential equation with homogeneous Dirichlet boundary conditions. It is well-known that local solutions satisfy the celebrated Pontryagin maximum principle. In this note, we will investigate an optimization method that is based on the maximum principle. We prove that the discrepancy in the maximum principle vanishes along the resulting sequence of iterates. Numerical experiments confirm the theoretical findings.

1. Introduction

In this note, we consider the following optimal control problem: Minimize

$$J(y,u) := \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \int_\Omega g(u(x))\,\mathrm{d}x \tag{1.1}$$

over all $u \in L^2(\Omega)$ and $y \in H^1_0(\Omega)$ satisfying

$$\begin{aligned} -\Delta y &= u && \text{in } \Omega,\\ y &= 0 && \text{on } \partial\Omega. \end{aligned}$$

Here, $\Omega \subset \mathbb{R}^d$ is a bounded domain, and $g : \mathbb{R} \to \bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ is assumed to be proper and lower semicontinuous. In addition, we require

$$\lim_{|v|\to\infty} \frac{g(v)}{|v|} = +\infty. \tag{1.2}$$

Note, that we assume neither continuity nor convexity of g. Hence, it is impossible to prove existence of solutions of (1.1). In fact, one can construct problems without solution, see [18, Section 4.5]. In this note, we will work with the example

$$g(u) := \frac{\alpha}{2}u^2 + I_{\mathbb{Z}}(u) = \begin{cases} \frac{\alpha}{2}u^2 & \text{if } u \in \mathbb{Z}\\ +\infty & \text{otherwise,} \end{cases} \tag{1.3}$$

where $\alpha > 0$. If g is assumed to be convex and continuous, then existence of solutions of (1.1) can be proven by the direct method of the calculus of variations [17]. Let us remark that by the above assumptions g is bounded from below.

If solutions exist, then the Pontryagin maximum principle [11] is a necessary optimality condition. Its main feature is that no differentiability with respect to the controls is needed, and so it is perfectly suited for the problems considered here. In fact, due to the structure of the problem (linear state equation, convexity of J with respect to y), the maximum principle is sufficient. We refer to [2,4,5,12] for the Pontryagin maximum principle applied to optimal control problems for partial differential equations. The goal of this note is to construct an algorithm to solve the maximum principle. We will comment on related work in Section 4.

2. Sensitivity analysis

In this section, we will perform a sensitivity analysis with respect to perturbations of the control with characteristic functions. The setup is as follows: Let $u, \tilde{u} \in L^2(\Omega)$ be feasible controls, i.e., the integrals $\int_\Omega g(u)\,\mathrm{d}x$ and $\int_\Omega g(\tilde{u})\,\mathrm{d}x$ exist. Let $B \subset \Omega$ be measurable. We define

$$u_B := u + \chi_B(\tilde{u} - u).$$

 ISBN 978-84-10135-30-7

Let y, y_B be the uniquely determined weak solutions of

$$\begin{aligned} -\Delta y = u \quad -\Delta y_B = u_B \quad &\text{in } \Omega, \\ y = 0 \qquad y_B = 0 \quad &\text{on } \partial\Omega. \end{aligned}$$

Let $p \in H_0^1(\Omega)$ be the weak solution of the adjoint equation

$$\begin{aligned} -\Delta p = y - y_d \quad &\text{in } \Omega, \\ p = 0 \quad &\text{on } \partial\Omega. \end{aligned}$$

The goal is now to estimate $J(y_B, u_B) - J(y, u)$ in terms of $u, \tilde{u}, p$ and the Lebesgue measure $|B|$ of B. Here, we have the following result.

Lemma 2.1 *Under the assumptions above, we have*

$$J(y_B, u_B) - J(y, u) = \int_B (\tilde{u} - u)p + g(\tilde{u}) - g(u)\,\mathrm{d}x + \frac{1}{2}\|y_B - y\|_{L^2(\Omega)}^2.$$

Proof This follows directly from the definition of p and u_B:

$$\begin{aligned} J(y_B, u_B) - J(y, u) &= \frac{1}{2}\|y_B - y_d\|_{L^2(\Omega)}^2 + \int_\Omega g(u_B)\,\mathrm{d}x - \frac{1}{2}\|y - y_d\|_{L^2(\Omega)}^2 - \int_\Omega g(u)\,\mathrm{d}x \\ &= \int_\Omega (y_B - y)(y - y_d) + \frac{1}{2}(y_B - y)^2\,\mathrm{d}x + \int_B g(\tilde{u}) - g(u)\,\mathrm{d}x \\ &= \int_B (\tilde{u} - u)p + g(\tilde{u}) - g(u)\,\mathrm{d}x + \frac{1}{2}\|y_B - y\|_{L^2(\Omega)}^2. \end{aligned}$$

□

We will now prove that $\|y_B - y\|_{L^2(\Omega)}^2$ is of higher order with respect to the Lebesgue measure $|B|$ of B.

Lemma 2.2 *There are constants $c > 0$ and $\nu > 1/2$ independent of $u, \tilde{u}, B$ such that*

$$\|y_B - y\|_{L^2(\Omega)} \le c\,|B|^\nu \cdot \|\tilde{u} - u\|_{L^\infty(\Omega)},$$

where $|B|$ denotes the Lebesgue measure of B. The constant ν can be chosen as

$$\nu = \begin{cases} 1 & \text{if } d \le 3, \\ 1 - \epsilon & \text{if } d = 4 \text{ for } \epsilon > 0, \\ \frac{1}{2} + \frac{2}{d} & \text{if } d > 4. \end{cases}$$

Proof We prove the claim by a well-known duality argument. Assume $d \le 3$. Let $w \in L^2(\Omega)$ be given. Let $z, q \in H_0^1(\Omega)$ be the weak solutions of

$$\begin{aligned} -\Delta z = w \quad -\Delta q = z \quad &\text{in } \Omega, \\ z = 0 \qquad q = 0 \quad &\text{on } \partial\Omega. \end{aligned}$$

Due to [15], there is $c > 0$ independent of w, z such that

$$\|z\|_{L^\infty(\Omega)} \le c\|w\|_{L^2(\Omega)}.$$

Testing the weak formulations with z and q yields

$$\|z\|_{L^2(\Omega)}^2 = \int_\Omega wq\,\mathrm{d}x \le \|w\|_{L^1(\Omega)}\|q\|_{L^\infty(\Omega)} \le c\|w\|_{L^1(\Omega)}\|z\|_{L^2(\Omega)}.$$

This proves $\|z\|_{L^2(\Omega)} \le c\|w\|_{L^1(\Omega)}$. Applying this estimate to $z := y_B - y$ and $w := u_B - u$ yields the claim with

$$\|y_B - y\|_{L^2(\Omega)} \le c\|u_B - u\|_{L^1(\Omega)} \le c\,|B| \cdot \|\tilde{u} - u\|_{L^\infty(\Omega)}.$$

In case $d > 3$ one can use the estimates from [3, Theorem 18]. □

Combining these results proves the following theorem.

Theorem 2.3 *Let $u, \tilde{u} \in L^\infty(\Omega)$. Let $B \subset \Omega$ be measurable. Let $\tilde{u}, y_B, y, p$ be defined as above. Then there are $\gamma > 0$ and $c > 0$ independent of $u, \tilde{u}, B$ such that*

$$J(y_B, u_B) - J(y, u) \le \int_B (\tilde{u} - u)p + g(\tilde{u}) - g(u)\,\mathrm{d}x + c\,|B|^{1+\gamma}\|\tilde{u} - u\|_{L^\infty(\Omega)}^2.$$

3. Pontryagin maximum principle

With the help of Theorem 2.3 we can prove the Pontryagin maximum principle.

Theorem 3.1 *Let $\bar{u} \in L^\infty(\Omega)$ be locally optimal with respect to $L^1(\Omega)$ topology for the control problem* (1.1)*. Let $\bar{y}, \bar{p} \in H_0^1(\Omega)$ be the optimal state and adjoint solving*

$$\begin{aligned} -\Delta \bar{y} = \bar{u} \quad -\Delta \bar{p} = \bar{y} - y_d \quad & \text{in } \Omega, \\ \bar{y} = 0 \qquad \bar{p} = 0 \qquad & \text{on } \partial\Omega. \end{aligned}$$

Let $v \in \mathbb{R}$ be such that $g(v) < +\infty$. Then

$$\bar{u}(x)\bar{p}(x) + g(\bar{u}(x)) \leq v\bar{p}(x) + g(v) \text{ for almost all } x \in \Omega. \tag{3.1}$$

Proof Let $v \in \mathbb{R}$ be such that $g(v) < +\infty$. Applying Theorem 2.3 with $u := \bar{u}, \tilde{u} := v$ yields

$$0 \leq J(y_B, u_B) - J(\bar{y}, \bar{u}) = \int_B (v - \bar{u})\bar{p} + g(\tilde{u}) - g(\bar{u})\,\mathrm{d}x + o(|B|).$$

By standard arguments based on the Lebesgue differentiation theorem, see, e.g., [10, Theorem 2.1], the claim follows. □

The maximum principle is a sufficient condition for the problem considered here.

Corollary 3.2 *Let $\bar{u} \in L^2(\Omega)$ satisfy the conclusion* (3.1) *of Theorem 3.1. Then $\bar{u}$ is global optimal for* (1.1)*.*

Proof Let $\tilde{u} \in L^2(\Omega)$ be an admissible control with associated state $\tilde{y}$. Then Lemma 2.1 with $B = \Omega$ yields

$$J(\tilde{y}, \tilde{u}) - J(\bar{y}, \bar{u}) = \int_\Omega (\tilde{u} - \bar{u})\bar{p} + g(\tilde{u}) - g(\bar{u})\,\mathrm{d}x + \frac{1}{2}\|\tilde{y} - \bar{y}\|^2_{L^2(\Omega)}.$$

Since $\bar{u}$ satisfies (3.1), the first expression is non-negative, which implies $J(\tilde{y}, \tilde{u}) - J(\bar{y}, \bar{u}) \geq \frac{1}{2}\|\tilde{y} - \bar{y}\|^2_{L^2(\Omega)} \geq 0$. □

4. Construction of an algorithm

We will now apply Theorem 2.3 with $u := u_k$ and $\tilde{u} := \tilde{u}_k$, where u_k is the current iterate of the algorithm to be devised. Let y_k and p_k be the associated state and adjoint. The control $\tilde{u}_k$ has to be computed in each iteration. Let B_k be measurable. Then we have

$$J(y_{B_k}, u_{B_k}) - J(y_k, u_k) = \int_{B_k} (\tilde{u}_k - u_k)p_k + g(\tilde{u}) - g(u_k)\,\mathrm{d}x + o(|B_k|). \tag{4.1}$$

The idea is now to choose $\tilde{u}_k$ and B_k such that $J(y_{B_k}, u_{B_k}) - J(y_k, u_k)$ is negative and to define the new iterate by

$$u_{k+1} = u_k + \chi_{B_k}(\tilde{u}_k - u_k).$$

In view of the maximum principle, Theorem 3.1, it is natural to choose $\tilde{u}_k$ as a function satisfying

$$\tilde{u}_k(x) \in \operatorname*{arg\,min}_{v \in \mathbb{R}} vp_k + g(v). \tag{4.2}$$

In addition, B_k will be chosen to get sufficient descent.

Let us comment on related work. The classic algorithm of [8] chooses $B_k := \Omega$, resulting in a fixed-point scheme to solve the maximum principle. The min-h method of [7] uses the update $u_{k+1} := u_k + t(\tilde{u}_k - u_k)$ with $t \in (0, 1]$, and is thus only suited for convex functions g. In the monograph [14], a method similar to ours is presented to solve optimal control problems with ODEs. Let us also also mention the review papers [6, 16]. In [9] binary control problems are solved with a similar approach: there a trust-region globalization is proposed, whereas we use an Armijo line-search to globalize.

As motivated above, we will compute $\tilde{u}_k$ as a result of the pointwise minimization

$$\tilde{u}_k(x) \in \operatorname*{arg\,min}_{v \in \mathbb{R}} vp_k + g(v).$$

Due to (1.2) this problem is solvable for all x. A measurable selection of this argmin-map exists [1]. For the example of g proposed in (1.3), we get

$$\tilde{u}_k(x) \in \operatorname{round}\left(-\frac{1}{\alpha}p_k(x)\right).$$

It remains to describe how B_k is chosen. Here, we are faced with two competing goals: In order to make the first term in (4.1) as small as possible, B_k has to be chosen as large as possible. However, to control the remainder term in (4.1), $|B_k|$ has to be chosen sufficiently small.

We propose the following line-search. Given $t \in (0, 1]$, choose B_t such that

$$\begin{aligned}\int_{B_t} (\tilde{u} - u_k)p_k + g(\tilde{u}) - g(u_k)\,\mathrm{d}x \le t \int_\Omega (\tilde{u} - u_k)p_k + g(\tilde{u}) - g(u_k)\,\mathrm{d}x,\\ |B_t| \le t \cdot |\Omega|.\end{aligned} \tag{4.3}$$

Due to the celebrated Lyapunov convexity theorem, see, e.g., [13, Theorem 5.5], a measurable set B_t satisfying (4.3) exists. Given t and B_t, we set $u_t := u_k + \chi_{B_t}(\tilde{u}_k - u_k)$. Let y_t be the associated state.

The parameter t_k is determined by the following procedure: Let t_k be the largest number in $\{\beta^l : l \in \mathbb{N} \cup \{0\}\}$, where $\beta \in (0, 1)$, that satisfies the descent condition

$$J(y_t, u_t) - J(y_k, u_k) \le \sigma \int_{B_t} (\tilde{u} - u_k)p_k + g(\tilde{u}) - g(u_k)\,\mathrm{d}x \tag{4.4}$$

where $\sigma \in (0, 1)$, and B_t is a measurable set satisfying (4.3). This condition is inspired by the well-known Armijo line-search in nonlinear optimization. If u_k does not satisfy the maximum principle, there is an admissible step-size t_k, and the resulting algorithm produces a new iterate with smaller value of the objective.

Lemma 4.1 *Suppose that*

$$\int_\Omega (\tilde{u} - u_k)p_k + g(\tilde{u}) - g(u_k)\,\mathrm{d}x < 0.$$

There is $t_0 > 0$ such that for all $t \in (0, t_0)$ condition (4.4) is satisfied.

Proof Due to Theorem 2.3, we have

$$\begin{aligned}J(y_t, u_t) - J(y_k, u_k) - \sigma \int_{B_t} (\tilde{u} - u_k)p_k + g(\tilde{u}) - g(u_k)\,\mathrm{d}x\\ \le (1 - \sigma) \int_{B_t} (\tilde{u} - u_k)p_k + g(\tilde{u}) - g(u_k)\,\mathrm{d}x + o(t)\\ \le t(1 - \sigma) \int_\Omega (\tilde{u} - u_k)p_k + g(\tilde{u}) - g(u_k)\,\mathrm{d}x + o(t),\end{aligned}$$

which proves the claim. □

The resulting algorithm is sketched in Algorithm 1.

Let us now turn to the convergence analysis of Algorithm 1. Here, we follow the related analysis in [19]. Let us define

$$\rho_k := \int_\Omega (\tilde{u}_k - u_k)p_k + g(\tilde{u}) - g(u_k)\,\mathrm{d}x.$$

Due to the choice of $\tilde{u}_k$ in (4.2), it follows $\rho_k \le 0$. If $\rho_k = 0$ then u_k satisfies the maximum principle Theorem 3.1, and the corresponding control u_k is optimal by Corollary 3.2.

Lemma 4.2 *Let (u_k) be an infinite sequence generated by Algorithm 1. Then*

$$\sum_{k=0}^{\infty} t_k \|\rho_k\|_{L^1(\Omega)} < +\infty.$$

Algorithm 1 Maximum-principle based descent algorithm

Choose $\beta \in (0,1)$, $\sigma \in (0,1)$, u_0 with $\int_\Omega g(u_0)\,\mathrm{d}x < \infty$, $\delta_{\mathrm{tol}} \geq 0$. Set $k := 0$.
loop ▷ Gradient descent
  Compute state y_k and adjoint p_k associated to u_k.
  Compute $\tilde{u}_k$ as in (4.2).
  if $\left|\int_\Omega (\tilde{u}_k - u_k)p_k + g(\tilde{u}) - g(u_k)\,\mathrm{d}x\right| \leq \delta_{\mathrm{tol}}$ **then** ▷ Termination criterion
    return u_k
  end if
  $t := 1$.
  loop ▷ Armijo line-search
    Compute $B_{k,t}$ satisfying (4.3).
    Compute $J(y_t, u_t)$.
    if (4.4) is satisfied **then**
      break
    end if
    $t := \beta \cdot t$.
  end loop
  $t_k := t$. ▷ Update
  $u_{k+1} := u_k + \chi_{B_{k,t_k}}(\tilde{u}_k - u_k)$.
  $k := k + 1$.
end loop

Proof Using conditions (4.4) and (4.3) shows

$$J(y_{k+1},u_{k+1})-J(y_k,u_k) \leq \sigma \int_{B_{t_k}} (\tilde{u}-u_k)p_k+g(\tilde{u})-g(u_k)\,\mathrm{d}x \leq t_k \int_\Omega (\tilde{u}-u_k)p_k+g(\tilde{u})-g(u_k)\,\mathrm{d}x = -t_k\|\rho_k\|_{L^1(\Omega)}.$$

Due to (1.2), g has a global minimum and is bounded from below, so that J is bounded from below by some $M \in \mathbb{R}$. Summing this inequality over $k \in \mathbb{N}$ and using $J \geq M$ proves $\sum_{k=1}^{\infty} t_k\|\rho_k\|_{L^1(\Omega)} \leq J(y_0,u_0) - M < \infty$. □

For simplicity, we assume for the subsequent convergence analysis that

$$\operatorname{dom} g := \{v : \ g(v) < \infty\} \tag{4.5}$$

is compact. Then the set of iterates (u_k) and $(\tilde{u}_k)$ is uniformly bounded in $L^\infty(\Omega)$.

Corollary 4.3 *Assume* (4.5). *Let* $M > 0$ *such that* $\operatorname{dom} g \subset [-M, +M]$. *Then* $\|u_k\|_{L^\infty(\Omega)} \leq M$ *and* $\|\tilde{u}_k\|_{L^\infty(\Omega)} \leq M$ *for all* k.

Theorem 4.4 *Assume* (4.5). *Either the Algorithm 1 stops after finitely many steps with*

$$\left|\int_\Omega (\tilde{u}_k - u_k)p_k + g(\tilde{u}) - g(u_k)\,\mathrm{d}x\right| \leq \delta_{\mathrm{tol}}$$

(so that u_k *satisfies the maximum principle if* $\delta_{\mathrm{tol}} = 0$*), or*

$$\int_\Omega (\tilde{u} - u_k)p_k + g(\tilde{u}) - g(u_k)\,\mathrm{d}x \to 0,$$

i.e., the residual in the maximum principle tends to zero, and (u_k) *is a minimizing sequence.*

Proof We follow the proof of the related result [19, Theorem 6.7]. Let us assume the algorithm generates an infinite sequence of iterates. Let k be such that $t_k < 1$. Due to the line-search procedure of Algorithm 1, it follows that $t := \beta^{-1}t_k \leq 1$ violates the descent condition (4.4), that is

$$0 < J(y_t,u_t) - J(y_k,u_k) - \sigma \int_{B_t} (\tilde{u} - u_k)p_k + g(\tilde{u}) - g(u_k)\,\mathrm{d}x.$$

As in the proof of Lemma 4.1, we get from Theorem 2.3

$$0 < t(1-\sigma)\int_\Omega (\tilde{u}-u_k)p_k + g(\tilde{u}) - g(u_k)\,\mathrm{d}x + c\,|t|^{1+\gamma}\|\tilde{u}-u\|_{L^\infty(\Omega)}.$$

Together with Corollary 4.3, we get

$$0 < -t(1-\sigma)\|\rho_k\|_{L^1(\Omega)} + c|t|^{1+\gamma},$$

where c is independent of k. This implies

$$\|\rho_k\|_{L^1(\Omega)} \le c t_k^\gamma$$

for all k such that $t_k < 1$. With Lemma 4.2, we get

$$+\infty > \sum_{k=0}^{\infty} t_k\|\rho_k\|_{L^1(\Omega)} = \left(\sum_{k:\,t_k=1}\|\rho_k\|_{L^1(\Omega)}\right) + \left(\sum_{k:\,t_k<1} t_k\|\rho_k\|_{L^1(\Omega)}\right) \ge \left(\sum_{k:\,t_k=1}\|\rho_k\|_{L^1(\Omega)}\right) + c\left(\sum_{k:\,t_k<1}\|\rho_k\|_{L^1(\Omega)}^{1+\frac{1}{\gamma}}\right),$$

which results in $\lim_{k\to\infty}\|\rho_k\|_{L^1(\Omega)} = 0$. Hence, the algorithm stops after finitely many iterations if $\delta_{\mathrm{tol}} > 0$. □

5. Numerical results

Let us now report about numerical experiments with Algorithm 1. Here, we consider the optimal control problem

$$J(y,u) := \frac{1}{2}\|y-y_d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2}\|u\|_{L^2(\Omega)}^2 + I_{\mathbb{Z}\cap[-b,b]}(u)$$

over all $u \in L^2(\Omega)$ and $y \in H_0^1(\Omega)$ satisfying

$$\begin{aligned} -\Delta y &= u \quad \text{in } \Omega,\\ y &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

This fits into the setting of the paper with the choice

$$g(v) := \frac{\alpha}{2}v^2 + I_{\mathbb{Z}\cap[-b,b]}$$

Here, we chose $\Omega = (0,1)^2$,

$$y_d(x_1,x_2) = 10x_1\sin(5x_1)\cos(7x_2), \quad \alpha = 0.01, \quad \beta = 0.01, \quad b = 10.$$

We discretized the problem with piecewise linear finite elements on a regular mesh for state and adjoint variables, while the control was discretized with piecewise constant finite elements. We report the results for a sequence of different meshes, where the finest mesh has mesh-size $h = 1.41\cdot 10^{-3}$ resulting in $\approx 2\cdot 10^6$ degrees of freedom for the control variables, which results in a mixed-integer optimization problem with $\approx 2\cdot 10^6$ integer variables. In the implementation of Algorithm 1 a greedy strategy was used to determine B_t. The loop in Algorithm 1 was terminated if in the inner loop $t|\Omega|$ was smaller than any of the elements in the grid.

Now let us report about some of the results. The optimal control can be seen in the left plot of Figure 1. In the right plot, we report about the iteration history of the residual $\|\rho_k\|_{L^1(\Omega)}$. Surprisingly, the iterations seem to be mesh independent. In addition, for this particular problem a very small number of iterations was needed to optimize over $2\cdot 10^6$ discrete control variables.

This is underlined by the results in Table 1. It shows for different discretizations the final value of the objective J and the final value of the residual $\|\rho\|_{L^1(\Omega)}$. As can be seen from the last column of this table, very few outer iterations are needed. In conclusion, this new algorithm seems to be capable of solving quite challenging mixed-inter programs.

Acknowledgements

This research was supported by the German Research Foundation (DFG) under grant number WA 3626/3-2 within the priority program "Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization" (SPP 1962). The author thanks Anna Lentz for comments on an earlier version of this manuscript.

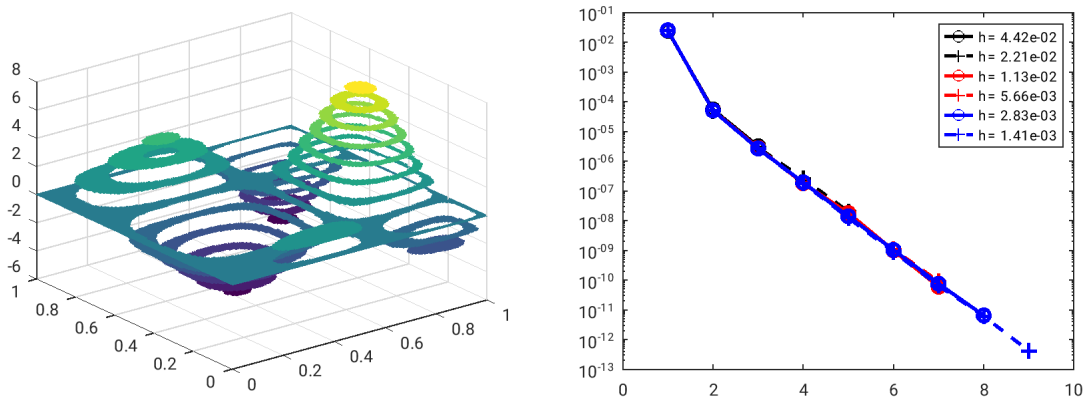


Fig. 1 Optimal control (left), iteration history (right)

h	J	$\|\rho\|_{L^1(\Omega)}$	It
$4.42 \cdot 10^{-2}$	4.706	$3.20 \cdot 10^{-6}$	4
$2.21 \cdot 10^{-2}$	5.048	$2.02 \cdot 10^{-8}$	6
$1.13 \cdot 10^{-2}$	5.210	$6.00 \cdot 10^{-11}$	8
$5.66 \cdot 10^{-3}$	5.293	$8.91 \cdot 10^{-11}$	8
$2.83 \cdot 10^{-3}$	5.334	$6.46 \cdot 10^{-12}$	9
$1.41 \cdot 10^{-3}$	5.354	$4.11 \cdot 10^{-13}$	10

Tab. 1 Iteration history

References

[1] J.-P. Aubin and H. Frankowska. *Set-Valued Analysis*, volume 2 of *Systems & Control: Foundations & Applications*. Birkhäuser Boston, Inc., Boston, MA, 1990. doi:10.1007/978-0-8176-4848-0.

[2] F. Bonnans and E. Casas. An extension of Pontryagin's principle for state-constrained optimal control of semilinear elliptic equations and variational inequalities. *SIAM J. Control Optim.*, 33(1):274–298, 1995. doi:10.1137/S0363012992237777.

[3] H. Brézis and W. A. Strauss. Semi-linear second-order elliptic equations in L^1. *J. Math. Soc. Japan*, 25:565–590, 1973. doi:10.2969/jmsj/02540565.

[4] E. Casas, J.-P. Raymond, and H. Zidani. Pontryagin's principle for local solutions of control problems with mixed control-state constraints. *SIAM J. Control Optim.*, 39(4):1182–1203, 2000. doi:10.1137/S0363012998345627.

[5] E. Casas. Pontryagin's principle for optimal control problems governed by semilinear elliptic equations. In *Control and estimation of distributed parameter systems: nonlinear phenomena (Vorau, 1993)*, volume 118 of *Internat. Ser. Numer. Math.*, pages 97–114. Birkhäuser, Basel, 1994. doi:10.1007/978-3-0348-8530-0_6.

[6] F. L. Chernous'ko and A. A. Lyubushin. Method of successive approximations for solution of optimal control problems. *Optimal Control Appl. Methods*, 3(2):101–114, 1982. doi:10.1002/oca.4660030201.

[7] R. G. Gottlieb. Rapid convergence to optimum solutions using a min-h strategy. *AIAA Journal*, 5(2):322–329, 1967. doi:10.2514/3.3960.

[8] I. A. Krylov and F. L. Černous'ko. The method of successive approximations for solving optimal control problems. *Ž. Vyčisl. Mat i Mat. Fiz.*, 2:1132–1139, 1962. (In Russian).

[9] P. Manns, M. Hahn, C. Kirches, S. Leyffer, and S. Sager. On convergence of binary trust-region steepest descent. *J. Nonsmooth Anal. Optim.*, 4, 2023. doi:10.46298/jnsao-2023-10164.

[10] C. Natemeyer and D. Wachsmuth. A proximal gradient method for control problems with non-smooth and non-convex control cost. *Comput. Optim. Appl.*, 80(2):639–677, 2021. doi:10.1007/s10589-021-00308-0.

[11] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko. *The mathematical theory of optimal processes*. Interscience Publishers John Wiley & Sons, Inc., New York-London, 1962. Translated from the Russian by K. N. Trirogoff; edited by L. W. Neustadt.

[12] J. P. Raymond and H. Zidani. Pontryagin's principle for state-constrained control problems governed by parabolic equations with unbounded controls. *SIAM J. Control Optim.*, 36(6):1853–1879, 1998. doi:10.1137/S0363012996302470.

[13] W. Rudin. *Functional analysis*. McGraw-Hill Series in Higher Mathematics. McGraw-Hill Book Co., New York-Düsseldorf-Johannesburg, 1973.

[14] V. A. Srochko. *Iterative methods for solving optimal control problems*. Fismatlit, Moscow, 2000. (In Russian).

[15] G. Stampacchia. Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus. *Ann. Inst. Fourier (Grenoble)*, 15(fasc. 1):189–258, 1965. URL: http://www.numdam.org/item?id=AIF_1965__15_1_189_0.

[16] A. S. Strekalovsky. Modern methods for solving nonconvex optimal control problems. *The Bulletin of Irkutsk State University. Series Mathematics*, 8:141–163, 2014. (In Russian).

[17] F. Tröltzsch. *Optimal control of partial differential equations*, volume 112 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2010. Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels. doi:10.1090/gsm/112.

[18] D. Wachsmuth. Iterative hard-thresholding applied to optimal control problems with $L^0(\Omega)$ control cost. *SIAM J. Control Optim.*, 57(2):854–879, 2019. doi:10.1137/18M1194602.

[19] D. Wachsmuth. A topological derivative-based algorithm to solve optimal control problems with $L^0(\Omega)$ control cost. *J. Nonsmooth Anal. Optim.*, 5, 2024. doi:10.46298/jnsao-2024-12366.

French-German-Spanish Conference on Optimization
Gijón, June 18-21, 2024
(pp. 116-123)

Progress and future directions in machine learning through control theory

Enrique Zuazua[1,2,3]

1. enrique.zuazua@fau.de Chair for Dynamics, Control, Machine Learning, and Numerics, Alexander von Humboldt-Professorship, Department of Mathematics, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Germany
2. Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Madrid, Spain
3. Chair of Computational Mathematics, Fundación Deusto. Av. de las Universidades, 24, 48007 Bilbao, Basque Country, Spain

Abstract

This paper presents our recent advancements at the intersection of machine learning and control theory. We focus specifically on utilizing control theoretical tools to elucidate the underlying mechanisms driving the success of machine learning algorithms. By enhancing the explainability of these algorithms, we aim to contribute to their ongoing improvement and more effective application. Our research explores several critical areas:

Firstly, we investigate the memorization, representation, classification, and approximation properties of residual neural networks (ResNets). By framing these tasks as simultaneous or ensemble control problems, we have developed nonlinear and constructive algorithms for training. Our work provides insights into the parameter complexity and computational requirements of ResNets.

Similarly, we delve into the properties of neural ODEs (NODEs). We demonstrate that autonomous NODEs of sufficient width can ensure approximate memorization properties. Furthermore, we prove that by allowing biases to be time-dependent, NODEs can track dynamic data. This showcases their potential for synthetic model generation and helps elucidate the success of methodologies such as Reservoir Computing.

Next, we analyze the optimal architectures of multilayer perceptrons (MLPs). Our findings offer guidelines for designing MLPs with minimal complexity, ensuring efficiency and effectiveness for supervised learning tasks.

The generalization and prediction capacity of trained networks plays a crucial role. To address these properties, we present two nonconvex optimization problems related to shallow neural networks, capturing the "sparsity" of parameters and robustness of representation. We introduce a "mean-field" model, proving, via representer theorems, the absence of a relaxation gap. This aids in designing an optimal tolerance strategy for robustness and, through convexification, efficient algorithms for training.

In the context of large language models (LLMs), we explore the integration of residual networks with self-attention layers for context capture. We treat "attention" as a dynamical system acting on a collection of points and characterize their asymptotic dynamics, identifying convergence towards special points called leaders. These theoretical insights have led to the development of an interpretable model for sentiment analysis of movie reviews, among other possible applications.

Lastly, we address federated learning, which enables multiple clients to collaboratively train models without sharing private data, thus addressing data collection and privacy challenges. We examine training efficiency, incentive mechanisms, and privacy concerns within this framework, proposing solutions to enhance the effectiveness and security of federated learning methods.

Our work underscores the potential of applying control theory principles to improve machine learning models, resulting in more interpretable and efficient algorithms. This interdisciplinary approach opens up a fertile ground for future research, raising profound mathematical questions and application-oriented challenges and opportunities.

1. Introduction

The impact of machine learning (ML) and artificial intelligence (AI) in science is leading to rich and innovative lines of research in applied mathematics. There is a significant need for theoretical foundations that ensure the performance, reliability, and interpretability of ML methods. Specifically, mathematical models are required to understand and optimize rapidly emerging computational architectures. This challenge can be addressed through the lens of control theory, a combination that offers great potential.

In this paper, we discuss recent results from our group that explore the application of control tools to some of the main architectures and methods in ML, namely neural networks, self-attention mechanisms, and federated learning.

Control theory lies at the foundation of ML [15]. Aristotle anticipated control theory when he described the need for automated processes to free humans from their heaviest tasks [4]. In the 1940s, Norbert Wiener

redefined the term "cybernetics," previously coined by André-Marie Ampère, as "the science of communication and control in animals and machines," which reflected the discipline's definitive contribution to the industrial revolution.

Wiener's definition involves two conceptual binomials. The first is control-communication: the need for quality information about the state of the system to make the right decisions, reach given objectives, and avoid risky regimes. The second binomial is animal-machine: as anticipated by Aristotle, humans aim to build machines to perform routine tasks. These concepts are integral to contemporary ML. The close link between control theory and ML, and more generally AI, is thus inherent in Wiener's definition. Once more, we stand on the shoulders of giants.

2. Control-based supervised learning via neural networks

Supervised learning is one of the main paradigms of machine learning (ML), aiming to define a map that approximates an unknown function $f : \mathcal{X} \to \mathcal{Y}$ using a training dataset $\{(x_i, y_i)\}_{i=1}^N$. Neural networks form a widely used class of functions to approximate f, and among these, residual networks have proven to be particularly effective. In the continuous-time limit, these discrete systems, like for instance Residual Neural Networks (ResNets),

$$x^{k+1} = x^k + W_k \boldsymbol{\sigma}(A^k x_i^k + b^k), \quad k \in [L], \tag{2.1}$$

transform into the so-called Neural ODE (NODE):

$$\begin{cases} \dot{x}(t) = W(t)\boldsymbol{\sigma}(A(t)x(t) + b(t)), & t \in (0,T), \\ x(0) = x_i, \end{cases} \tag{2.2}$$

for all $i \in [N] := \{1, \dots, N\}$. Here, $x = x(t)$ is the state if the system, representing the data under consideration, evolving continuously on time in the ambient space, $(W(t), A(t), b(t)) \in \Theta_p := L^\infty\big((0,T); \mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d} \times \mathbb{R}^p\big)$ are piecewise constant controls with L discontinuities (which play the role of the NN parameters to be trained), $L, p \geq 1$ represent the depth and the width of the model, respectively, and $\boldsymbol{\sigma} : \mathbb{R}^p \to \mathbb{R}^p$ is a Lipschitz-continuous non-linearity defined component-wise, a common example being the *rectified linear unit* (ReLU): $x \mapsto \max\{x, 0\}$.

One of the main advantages of NODEs is the possibility to reinterpret several machine learning paradigms using tools from differential equations and their control. For example, data classification can be formulated as a simultaneous control problem for (2.2), the goal being to build controls (W, A, b) driving all initial data $\{x_i\}_{i=1}^N$ to their corresponding targets $\{y_i\}_{i=1}^N$ (prescribed according to the labels) through the flow map generated by (2.2).

In [11], we prove the simultaneous controllability of (2.2) for the single-neuron width case ($p = 1$) via an inductive algorithm that constructs explicit, piecewise constant controls (W, A, b) to sequentially guide each point x_i to its target y_i. Moreover, using similar techniques, we obtain a result of universal approximation in $\|\cdot\|_{L^2}$ for NODEs. Below, we state the two main results from [11]:

Theorem 2.1 (Controllability) *Let $N \geq 1$, $d \geq 2$, and $T > 0$. Consider any dataset $\{x_i, y_i\}_{i=1}^N \subset \mathbb{R}^d$ with $x_i \neq x_j$ and $y_i \neq y_j$ for $i \neq j$. Then, there exists a piecewise constant control $(W, A, b) \in \Theta_1$ (with $p = 1$) such that the flow map Φ_T generated by* (2.2) *satisfies*

$$\Phi_T(x_i) = y_i, \qquad \text{for all } i = 1, \dots, N.$$

Furthermore, the depth of the model is $L = 3N$.

Theorem 2.2 (Approximation) *Let $d \geq 2$, $T > 0$ and a bounded set $\Omega \subset \mathbb{R}^d$. Then, for any $f \in L^2(\Omega; \mathbb{R}^d)$ and $\varepsilon > 0$ there exists a piecewise constant control $(W, A, b) \in \Theta_1$ (with $p = 1$) such that the flow map Φ_T generated by* (2.2) *satisfies*

$$\|\Phi_T - f\|_{L^2(\Omega)} < \varepsilon.$$

The simultaneous control result in theorem 2.1 and its proof opens paths for new methodologies in data classification, albeit requiring very high complexity (it scales with N). In [3], we reduce the complexity of the controls for binary classification by proposing new algorithms based on predetermined point clusterings. Our strategy aims to probabilistically reduce the number of parameters needed by leveraging the spatial structure of the data distribution, assuming that the points are in general position, i.e., no $d + 1$ points can lie on the same hyperplane in $\mathbb{R}^d$, which is generically fulfilled by random datasets.

Theorem 2.3 *Let $d \geq 2$ and $N \geq 1$. For any dataset $\{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \{1, 0\}$ in general position and any $j \in \{1, \dots, d\}$, there exist $T > 0$ and a piecewise constant control $(W, A, b) \in \Theta_1$ (with $p = 1$) with $L = 4\lceil m/d \rceil - 1$ discontinuities, where $m = \min(\#\{i : y_i = 1\}, \#\{i : y_i = 0\})$, such that the flow map generated by* (2.2) *satisfies*

$$\Phi_T(x_i)^{(j)} < 1 \quad \text{if } y_i = 1 \quad \text{and} \quad \Phi_T(x_i)^{(j)} > 1 \quad \text{if } y_i = 0, \qquad \text{for all } i = 1, \dots, N.$$

The described results are focused on the simplified version of (2.2) with $p = 1$ neurons per layer. In [2], we focus on the role played by the architecture through the interplay between the depth L and width p. Our findings reveal a balancing trade-off, as shown in the following result:

Theorem 2.4 *Let $N \geq 1$, $d \geq 2$, $T > 0$. Consider any dataset $\{x_i, y_i\}_{i=1}^N \subset \mathbb{R}^d$ with $x_i \neq x_j$ and $y_i \neq y_j$ if $i \neq j$. For any $p \geq 1$, there exists a piecewise constant control $(W, A, b) \in \Theta_p$ such that the flow map Φ_T generated by* (2.2) *satisfies*

$$\Phi_T(x_i) = y_i, \qquad \text{for all } i = 1, \dots, N.$$

Furthermore, the depth of the model is $L = 2\lceil N/p \rceil$.

In the wide limit, where $L = 0$, the system (2.2) becomes autonomous and a separate study is required. We address the relaxed problem of ε-approximate controllability of N pairs of points and establish an explicit error decay by uniformly approximating a custom-built Lipschitz vector field that effectively interpolates the dataset:

Theorem 2.5 *Let $N \geq 1$, $d \geq 2$ and $T > 0$ be fixed. Consider any dataset $\{x_i, y_i\}_{i=1}^N \subset \mathbb{R}^d$ with $x_i \neq x_j$. For each $p \geq 1$, there exists a control $(W, A, b) \in \Theta_p$ such that the flow map Φ_T generated by* (2.2) *satisfies*

$$\sup_{i=1,\dots,N} |y_i - \Phi_T(x_i)| \leq C \frac{\log_2(\kappa)}{\kappa^{1/d}},$$

where $\kappa = (d + 2)dp$ is the number of neurons in the model, and $C > 0$ is a constant depending on d, T, but independent of κ.

The study of the autonomous system is closely related to the turnpike principle paradigm, coined by John von Neumann, which ensures that optimal control strategies remain almost steady over long time periods. In [5], we have analyzed the implications of this principle for designing simplified and more stable architectures for deep ResNets.

An extension of the developed theory reformulates the continuous model in terms of transport equations, through the classical link between (2.2), seen as the ODE of characteristics, and the hyperbolic transport PDE, leading to the following neural transport model:

$$\partial_t \rho + \operatorname{div}_x(W(t)\boldsymbol{\sigma}(A(t)x + b(t))\rho) = 0. \tag{2.3}$$

Transforming one given probability measure into another, up to an arbitrarily small Wasserstein-1 error [2, 11] or total variation error [12], can be reinterpreted as a control problem for (2.3). The first approach allows us to build a bridge with the theory of optimal transport, whereas the latter, whose theorem statement we formulate below, has applications in generative modeling via the technique known as normalizing flows.

Theorem 2.6 *Given two probability densities $\rho_0, \rho_T \in L^1(\mathbb{R}^d)$, for any $T > 0$ and for all $\varepsilon > 0$, there exist piecewise constant controls $(w, a, b) \in \Theta_1$ such that the solution of* (2.3) *satisfies*

$$\|\rho(T) - \rho_T\|_{L^1(\mathbb{R}^d)} < \varepsilon.$$

In addition to ResNets and NODEs, we have analyzed the so-called multilayer perceptron deep NN:

$$x^{k+1} = \boldsymbol{\sigma}_{k+1}(A^k x^k + b^k), \quad k \in [L], \tag{2.4}$$

where $x^k \in \mathbb{R}^{d_k}$ denotes the state at layer/step $k \geq 1$, $A^k \in \mathbb{R}^{d_{k+1} \times d_k}$, $b^k \in \mathbb{R}^{d_{k+1}}$, and $\{d_k\}_{k=1}^L$ is a sequence of positive integers determining the dimension of the state and the width of (2.4) at the layer k. Here, $\boldsymbol{\sigma}_{k+1} : \mathbb{R}^{d_{k+1}} \to \mathbb{R}^{d_{k+1}}$ denotes the (component-wise) ReLU function, and $\max_k\{d_k\}$ the total width of (2.4). In [6], for a dataset of N elements in $\mathbb{R}^d, d \geq 1$, and M classes, we prove that (2.4) is simultaneously controllable with width 2 and at most $2N + 4M - 1$ layers. This is proven using an inductive algorithm that provides explicit values for the parameters. This result is sharp in the sense that (2.4) with width 1 cannot achieve simultaneous controllability. Additionally, in [6], the universal approximation (UA) for $L^p(\Omega; \mathbb{R}_+)$ functions (for $p \in [1, \infty)$ and $\Omega \subset \mathbb{R}^d$ bounded) is proven, using (2.4) with width $d + 1$, together with explicit convergence rates for $W^{1,p}$ functions, which can be extended to changing-sign functions too.

3. Representer theorem for shallow neural networks: sparsity and generalization

Besides NODEs, ResNets and deep NNs, we have also analysed the representational and generalization capacity of shallow NN, as conducted in [9]. The shallow NN is expressed as:

$$f_{\text{shallow}}(x,\Theta) \coloneqq \sum_{j=1}^{P} \omega_j \sigma(\langle a_j, x\rangle + b_j), \tag{3.1}$$

where $\Theta = \{(\omega_j, a_j, b_j) \in \mathbb{R}\times\Omega\}_{j=1}^{P}$, P denotes its width, and Ω is a compact subset of $\mathbb{R}^d$ containing a neighborhood of 0. We first investigate the representational capacity of (3.1).

Theorem 3.1 *Assume that σ is continuous and $\sigma(x) = 0$ for $x \le 0$ and $\sigma(x) > 0$ for $x > 0$. Fix any consistent dataset $\{(x_i, y_i) \in \mathbb{R}^{d+1}\}_{i=1}^{N}$. If $P \ge N$, then there exists $\Theta \in (\mathbb{R}\times\Omega)^P$ such that*

$$f_{\text{shallow}}(x_i,\Theta) = y_i, \quad \text{for all } i = 1,\dots,N.$$

For a fixed dataset $\{(x_i, y_i) \in \mathbb{R}^{d+1}\}_{i=1}^{N}$, Theorem 3.1 shows the existence of parameters for its exact representation by (3.1), $P = N$ being sufficient. Next, we consider an optimization problem, where the objective is to minimize the ℓ_1 norm of the neuron weights:

$$\inf_{\{(\omega_j,a_j,b_j)\in\mathbb{R}\times\Omega\}_{j=1}^{N}} \sum_{j=1}^{N} |\omega_j|, \qquad \text{s.t.} \sum_{j=1}^{N} \omega_j \sigma(\langle a_j, x_i\rangle + b_j) = y_i, \quad \text{for all } i = 1,\dots,N. \tag{P_0}$$

When $\{y_i\}_{i=1}^{N}$ represent observed labels affected by some level of noise, it is more meaningful to consider the previous optimization problem under certain tolerance on the error of the prediction. This leads to the following optimization problem parameterized by $\epsilon \ge 0$:

$$\inf_{\{(\omega_j,a_j,b_j)\in\mathbb{R}\times\Omega\}_{j=1}^{N}} \sum_{j=1}^{N} |\omega_j|, \qquad \text{s.t.} \left|\sum_{j=1}^{N} \omega_j \sigma(\langle a_j, x_i\rangle + b_j) - y_i\right| \le \epsilon, \quad \text{for all } i = 1,\dots,N. \tag{P_ϵ}$$

Problems (P_0) and (P_ϵ) are non-convex due to the non-linearity of σ, which induces the lack of convexity in their feasible sets. To cure this lack of convexity we consider the following convex relaxation problems:

$$\inf_{\mu\in\mathcal{M}(\Omega)} \|\mu\|_{\text{TV}}, \quad \text{s.t.} \int_{\Omega} \sigma(\langle a, x_i\rangle + b) d\mu(a,b) = y_i, \quad \text{for all } i = 1,\dots,N; \tag{PR_0}$$

$$\inf_{\mu\in\mathcal{M}(\Omega)} \|\mu\|_{\text{TV}}, \quad \text{s.t.} \left|\int_{\Omega} \sigma(\langle a, x_i\rangle + b) d\mu(a,b) - y_i\right| \le \epsilon, \quad \text{for all } i = 1,\dots,N, \tag{PR_ϵ}$$

where $\mathcal{M}(\Omega)$ represents the space of Radon measures on Ω, and $\|\cdot\|_{\text{TV}}$ denotes the total variation norm. We demonstrate that there is no gap between the primal problems and the relaxed ones, and that the extreme points of the relaxed solution sets have an atomic structure.

Theorem 3.2 *Under the setting of Theorem 3.1, the solution sets of* (PR_0) *and* (PR_ϵ)*, denoted by* $S(\mathrm{PR}_0)$ *and* $S(\mathrm{PR}_\epsilon)$*, are non-empty, convex and compact in the weak-$*$ sense. Moreover,*

$$\text{val}(\mathrm{PR}_0) = \text{val}(\mathrm{P}_0), \qquad \text{Ext}(S(\mathrm{PR}_0)) \subseteq \left\{ \sum_{j=1}^{N} \omega_j \delta_{(a_j,b_j)} \,\middle|\, (\omega_j,a_j,b_j)_{j=1}^{N} \in S(\mathrm{P}_0) \right\}, \tag{3.2}$$

$$\text{val}(\mathrm{PR}_\epsilon) = \text{val}(\mathrm{P}_\epsilon), \qquad \text{Ext}(S(\mathrm{PR}_\epsilon)) \subseteq \left\{ \sum_{j=1}^{N} \omega_j \delta_{(a_j,b_j)} \,\middle|\, (\omega_j,a_j,b_j)_{j=1}^{N} \in (\mathrm{P}_\epsilon) \right\}, \tag{3.3}$$

where $\text{Ext}(S)$ *represents the set of all extreme points of* S.

To study the generalization capacity of the shallow NN, we consider some testing dataset $\{(X', Y')\} = \{(x_i', y_i') \in \mathbb{R}^{d+1}\}_{i=1}^{N'}$ with $N' \in \mathbb{N}_+$, which differs from the training one. The generalization quality is determined by the performance of this shallow NN on the testing set (X', Y'), which is assessed by comparing the actual values $\{y_i'\}_{i=1}^{N'}$ with the predictions $\{f_{\text{shallow}}(x_i', \Theta)\}_{i=1}^{N'}$. Rather than evaluating differences individually, we analyze the discrepancies in their overall distributions to simplify the analysis. Let us denote by

$$m_x = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}, \quad m_y = \frac{1}{N}\sum_{i=1}^{N} \delta_{y_i}, \quad \tilde{m}_y = \frac{1}{N}\sum_{i=1}^{N} \delta_{f_{\text{shallow}}(x_i,\Theta)};$$

$$m_x' = \frac{1}{N'}\sum_{i=1}^{N'} \delta_{x_i'}, \quad m_y' = \frac{1}{N'}\sum_{i=1}^{N'} \delta_{y_i'}, \quad \tilde{m}_y' = \frac{1}{N'}\sum_{i=1}^{N'} \delta_{f_{\text{shallow}}(x_i',\Theta)}.$$

Theorem 3.3 *Assume that σ is L-Lipschitz. Let Θ be a solution of (P_ϵ) for some $\epsilon \geq 0$. Then,*

$$d_{\mathrm{KR}}(m'_y, \bar{m}'_y) \leq d_{\mathrm{KR}}(m_y, m'_y) + \begin{cases} \epsilon + d_{\mathrm{KR}}(m_x, m'_x) LD \,\mathrm{val}(\mathrm{P}_\epsilon), & \text{if } 0 \leq \epsilon \leq \|Y\|_{\ell^\infty}, \\ \|Y\|_{\ell^\infty}, & \text{otherwise,} \end{cases}$$

where $D = \sup_{(a,b)\in\Omega} \|a\|$ and d_{KR} represents the Kantorovich–Rubinstein distance.

In view of Theorem 3.3, the problem of minimizing the right-hand-side upper bound with respect to ϵ arises:

$$\inf_{0\leq\epsilon\leq\|Y\|_{\ell^\infty}} \mathcal{U}(\epsilon) := \epsilon + d_{\mathrm{KR}}(m_x, m'_x) LD \,\mathrm{val}(\mathrm{P}_\epsilon). \tag{UB}$$

By employing the dual analysis of problems (P_ϵ) and (P_0), we obtain the first-order optimality condition of (UB) in the following theorem. Let us denote by c_ϵ (resp. C_ϵ) the minimum (resp. maximum) value of the ℓ^1 norm of the dual solutions of (PR_ϵ) for any $\epsilon \geq 0$.

Theorem 3.4 *Under the setting of Theorem 3.1, the solution set of problem* (UB)*, denoted by S*(UB)*, is non-empty. Moreover, the following holds:*

1. *If $d_{\mathrm{KR}}(m_x, m'_x) < (LDc_0)^{-1}$, then $S(\mathrm{UB}) = \{0\}$.*
2. *If $d_{\mathrm{KR}}(m_x, m'_x) \geq (LDc_0)^{-1}$, then $\epsilon \in S(\mathrm{UB})$ if and only if $1/d_{\mathrm{KR}}(m_x, m'_x) LD \in [c_\epsilon, C_\epsilon]$.*

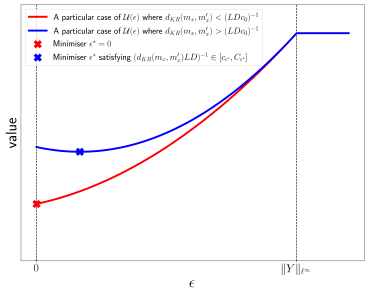


Fig. 3.1 The red and blue curves represent point 1 and 2 of Theorem 3.4, respectively. According to Theorem 3.4, when the distance between the training and testing sets is less than the threshold $(LDc_0)^{-1}$, it suffices to consider the exact representation problem (P_0). If $d_{\mathrm{KR}}(m_x, m'_x)$ exceeds this threshold, the optimal ϵ can be determined by solving the dual problem of (PR_ϵ).

4. Dynamical System Approximation via Semi-Autonomous NODEs

Going back to the NODE context, and with the aim of reducing their complexity, measured in terms of the number of switchings of the parameters, while preserving the exact representation capacity, in the upcoming work [7], we study NODEs of the form

$$\begin{cases} \dot{x} = W\boldsymbol{\sigma}(Ax + b(t)), & t \in (0,T), \\ x(0) = x_i, \end{cases} \tag{4.1}$$

where now the only time-dependent parameter is the bias $b = b(t)$. For this reason, we dub the model *Semi-Autonomous* NODE (SA-NODE), which is still non-autonomous, but with a complexity which is greatly reduced, since W and A are now time-independent. Theorem 2.1 continues to hold for (4.1) with no change in the hypotheses. Furthermore, the semi-autonomous structure appears naturally in the proof, as the time-dependency of the biases $b(t)$ is quickly seen to be necessary for tracking dynamic data, as the following result assures, [7].

Theorem 4.1 *Let $K \in \mathbb{R}^d$ be a fixed compact set and consider the non-autonomous ODE*

$$\begin{cases} \dot{\boldsymbol{z}}(t) = f(\boldsymbol{z}, t), \ t \in (0,T), \\ \boldsymbol{z}(0) = \boldsymbol{z}_0 \in K, \end{cases} \tag{4.2}$$

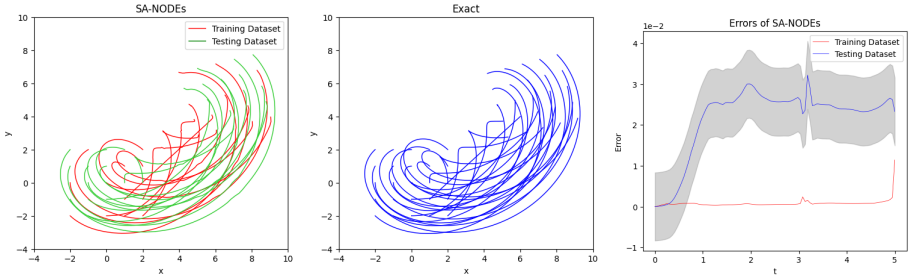
where $f : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$ is a continuous function and uniformly Lipschitz continuous in $\boldsymbol{z}$. For every $\varepsilon > 0$, there exist $p = p(\varepsilon)$, matrices $W \in \mathbb{R}^{d\times p}$, $A \in \mathbb{R}^{p\times d}$, and a function $b = b(t) \in L^\infty((0,T);\mathbb{R}^p)$ such that, for every $\boldsymbol{z}_0 \in K$, the solution $\boldsymbol{x} = \boldsymbol{x}(t)$ to the SA-NODE

$$\begin{cases} \dot{\boldsymbol{x}} = \displaystyle\sum_{i=1}^{p} w_i \sigma(a_i \cdot \boldsymbol{x} + b_i(t)), \\ \boldsymbol{x}(0) = \boldsymbol{z}_0, \end{cases} \tag{4.3}$$

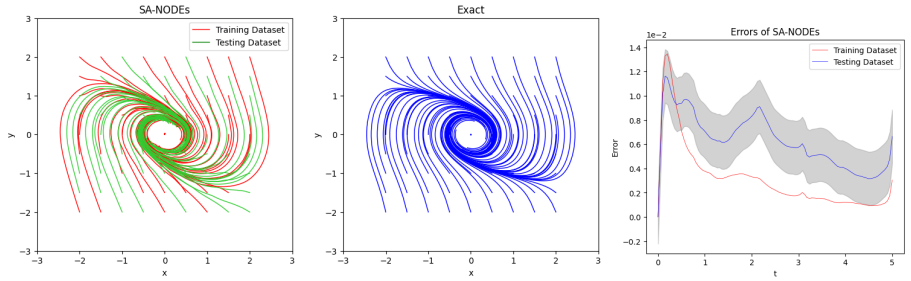
satisfies

$$\|\boldsymbol{z} - \boldsymbol{x}\|_{L^\infty([0,T];\mathbb{R}^d)} \leq \varepsilon. \tag{4.4}$$

In other words, SA-NODEs learn the *global* flow of the ODE, and not just the local information around one single trajectory or just the final profiles at $t = T$.



(a) SA-NODEs and exact solutions of linear ODEs for $\dot{z}_1 = t - z_2$, $\dot{z}_2 = z_1 - t$.



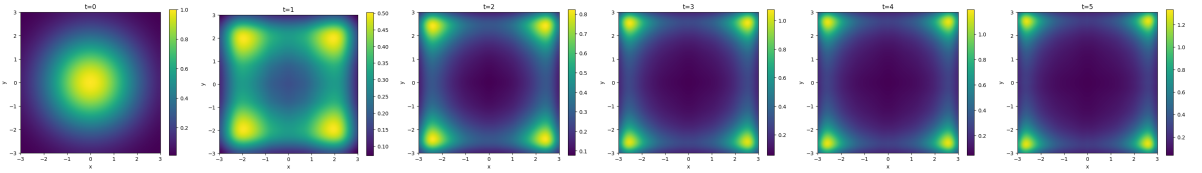
(b) SA-NODEs and exact solutions of nonlinear ODEs for $\dot{z}_1 = z_2$, $\dot{z}_2 = -z_1 - z_2^3$.

Fig. 4.1 SA-NODEs (left) and exact solutions (center) of linear and nonlinear ODEs. On the right, the mean and standard deviation bounds of the error $e(t)$, computed as the euclidean distance between the exact value of the trajectory and the predicted one.

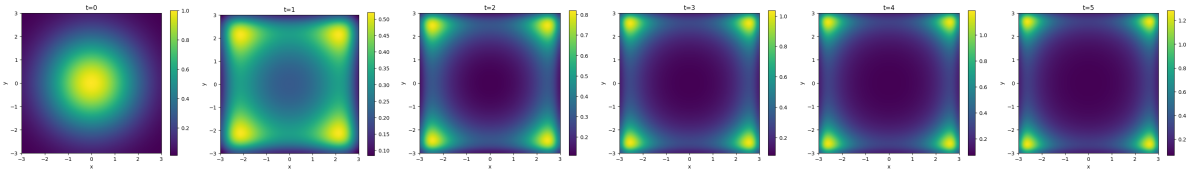
Notably, the semi-autonomous structure emerges spontaneously, roughly because Cybenko's universal approximation theorem yields an approximation of $f(\boldsymbol{z}, t)$ of the form

$$f(\boldsymbol{z}, t) \sim \sum_{i=1}^{p} w_i \sigma(a_i \cdot (\boldsymbol{z}, t)^\top + b_i) = \sum_{i=1}^{p} w_i \sigma(a_i \cdot \boldsymbol{z} + a_i^{d+1} t + b_i).$$

The SA-NODE structure arises naturally when renaming the term $a_i^{d+1} t + b_i$ as $b_i(t)$.



(a) SA-NODEs solutions



(b) Exact solutions

Fig. 4.2 SA-NODEs and exact solution of 2D transport equations $\rho_t + \mathrm{div}_x(f(x, y, t)\rho) = 0$, where the velocity field is $f(x, y, t) = \left[\sin(x)/(1 + t^2), \sin(y)/(1 + t^2)\right]^\top$. The initial datum is the gaussian profile $e^{-x^2-y^2}$.

Numerical results confirm that SA-NODEs are a promising architecture. They not only perform well on benchmark examples, such as linear and nonlinear dynamical systems (see Figures 4.1a-4.1b), but also on transport equations (as shown in Figure 4.2). In Figures 4.1a-4.1b, the simulated trajectories used for training are plotted in red. In contrast, the trajectories predicted from previously unseen initial data are plotted in green, demonstrating the excellent generalization properties of SA-NODEs.

Furthermore, SA-NODEs significantly outperform vanilla NODEs in terms of the number of epochs and neurons required to achieve suitable approximations of dynamical systems. On benchmark examples, and for a fixed number of epochs and neurons, SA-NODEs consistently achieve significantly smaller errors than vanilla NODEs, often by a couple of orders of magnitude. Additionally, even though the network widths are the same, SA-NODEs require less time to train than vanilla NODEs. This is because the number of parameters is reduced, with constant W and A. Consequently, by decreasing the number of parameters, SA-NODEs mitigate the tendency of vanilla NODEs to overfit. This showcases the potential of SA-NODEs for synthetic model generation and helps elucidate the success of methodologies such as Reservoir Computing.

5. Self-attention as a clustering mechanism and its role in LLMs

For supervised learning tasks in large language models (LLMs), capturing "context" or how words relate to one another in a sentence, is a key feature. The *transformer* is a state-of-the-art neural networks in LLMs, which builds on ResNets by alternating with *self-attention* layers exploiting the data structure. Heuristically, these layers capture the "context" at the sample level by mixing its rows based on similarity between them.

For this reason, the data samples used to train such models contain collections of words (i.e. sentences or paragraphs). More precisely, the training dataset is of the form $\{(Z_s, y_s)\}_{s=1}^N$, for matrices $Z_s \in \mathbb{R}^{n\times d}$, whose n rows encode words as points in Euclidean space $\mathbb{R}^d$.

For a fixed data sample $Z \in \mathbb{R}^{n\times d}$ with rows $z_1, \dots, z_n \in \mathbb{R}^d$, the (hardmax) self-attention model is given by

$$z_i^{k+1} = z_i^k + \frac{\alpha}{1+\alpha} \frac{1}{|\mathcal{C}_i(Z^k|} \sum_{j\in\mathcal{C}_i(Z^k} \left(z_j^k - z_i^k\right), \quad k \geq 0, \tag{5.1a}$$

where $z_i^0 = z_i$, Z^k contains the rows $z_1^k, \dots, z_n^k$, $A \in \mathbb{R}^{d\times d}$ is a symmetric positive definite matrix, $\alpha > 0$, and

$$\mathcal{C}_i(Z^k = \left\{j \in [n] \ : \ \left\langle Az_i^k, z_j^k \right\rangle = \max_{\ell\in[n]} \left\langle Az_i^k, z_\ell^k \right\rangle\right\}. \tag{5.1b}$$

In [1], we study the asymptotic behaviour of the self-attention dynamics (5.1) as $k \to \infty$. In particular, we prove that it exhibits clustering behaviour towards special points called *leaders*. As an application, we use our clustering results to design a simple and interpretable transformer-based model to solve the supervised learning task in LLMs of *sentiment analysis*. We use a benchmark dataset with movie reviews, labeled as positive or negative. The proposed model contains only three components with distinct roles: the encoder, mapping words to points in $\mathbb{R}^d$, whose role is to select meaningful words as leaders; our transformer (5.1), whose role is to capture "context" by clustering the majority of words towards the few most meaningful ones; and the decoder, whose role is to project the final point values to a real prediction by dividing $\mathbb{R}^d$ in two half-spaces and identifying each half-space with each sentiment. After training the model, our interpretation is verified with examples (cf. Figure 5.1).

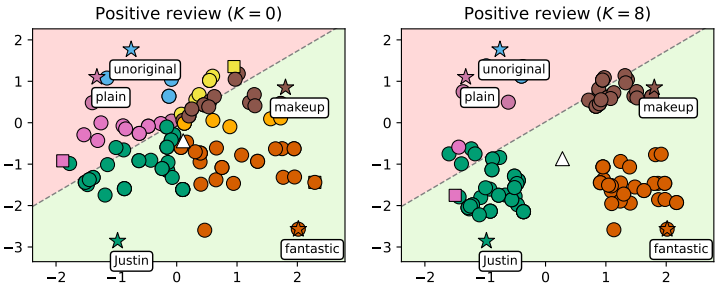


Fig. 5.1 Evolution of the encoded words of a positive review as they are processed by $K = 8$ transformer layers. Points are colored according to the point they follow, leaders are stars and tagged with the word they encode, squares are non-leaders who are followed by other points, circles are the remaining points, and the triangle is the mean word. The dashed line is the hyperplane separating the negative class (red) from the positive class (green).

6. Federated learning: training, incentive, and privacy

With the growing amount of distributed data, *federated learning* (FL) has emerged as a promising paradigm to address challenges like data collection and privacy protection in centralized learning approaches.

As in supervised learning, FL aims to learn a model to approximate $f : \mathcal{X} \to \mathcal{Y}$, but under the constraint that training data and labels are stored across distributed clients. Given m clients, the training of FL can be formulated as

$$\min_{\theta\in\mathcal{W}} \sum_{k=1}^m p_k \ell_k(\theta), \tag{6.1}$$

where $\theta \in \mathcal{W}$ are trainable parameters, $\ell_k : \mathcal{W} \to \mathbb{R}$ is client k's local loss function, commonly set as the empirical risk over its local dataset, and $p_k \geq 0$ with $\sum_{k=1}^{m} p_k = 1$ specifies the relative impact of client k.

To solve (6.1) efficiently, we propose in [14] an inexact and self-adaptive algorithm termed FedADMM-InSa. We design an inexactness criterion to guide each client to independently adjust its local training accuracy, leading to personalized training and better adaptation to heterogeneous data. Additionally, we present a self-adaptive scheme that dynamically adjusts each client's penalty parameter to enhance the robustness of our algorithm.

As in [14], existing research on FL primarily focuses on designing efficient learning algorithms. Most existing works do not consider that clients may be reluctant to engage without appropriate compensation (rewards from the server) for their training efforts. We address this issue in [8] by formulating incentive mechanisms in FL within a potential game framework. We investigate the uniqueness of the Nash equilibrium in these games and offer the server an easily calculable threshold for the reward, under which it can achieve effective incentives concerning clients' training efforts.

Moreover, the privacy benefits of FL (exchanging model parameters instead of data) can be compromised by data reconstruction attacks. In [13], we propose an approximate and weighted attack method to recover clients' private data under the widely used multiple-step local update scenarios. Experimental results validate the superiority of our attack method, emphasizing the need for effective defense mechanisms in FL to enhance privacy.

Acknowledgements

The author thanks all team members who contributed to develop the research summarised here and, in particular, Albert Alcalde, Antonio Álvarez-López, Martin Hernández, Ziqian Li, Kang Liu, Lorenzo Liverani, Yongcun Song and Ziqi Wang who contributed to draft this synthesis article.

Our research has been funded by the Alexander von Humboldt-Professorship program, ModConFlex Marie Curie Action, HORIZON-MSCA-2021-DN-01, COST Action MAT-DYN-NET, Transregio 154 Project "Mathematical Modelling, Simulation and Optimization Using the Example of Gas Networks" of the DFG, AFOSR-24IOE027, grants PID2020-112617GB-C22 and TED2021-131390B-I00 of MICINN (Spain).

For brevity we have only included references from our team. The articles below contain a much richer bibliography on each of the topics discussed.

References

[1] Albert Alcalde, Giovanni Fantuzzi, and Enrique Zuazua. Clustering in pure-attention hardmax transformers and its role in sentiment analysis. In preparation (2024).

[2] Antonio Álvarez-López, Arselane Hadj Slimane, and Enrique Zuazua. Interplay between depth and width for interpolation in neural ODEs. *arXiv preprint arXiv:2401.09902*, 2024.

[3] Antonio Álvarez-López, Rafael Orive-Illera, and Enrique Zuazua. Optimized classification with neural ODEs via separability. *arXiv preprint arXiv:2312.13807*, 2023.

[4] Enrique Fernández-Cara, and Enrique Zuazua. Control theory: History, mathematical achievements and perspectives. *Bol. Soc. Esp. Mat. Apl.*, 26, 79-140, 2003.

[5] Borjan Geshkovski and Enrique Zuazua. Turnpike in optimal control of PDEs, ResNets, and beyond. *Acta Numerica*, 31:135–263, 2022. Cambridge University Press.

[6] Martin Hernández and Enrique Zuazua. Deep neural networks: Multi-classification and universal approximation. In preparation (2024).

[7] Ziqian Li, Kang Liu, Lorenzo Liverani and Enrique Zuazua. Universal Approximation of Dynamical Systems by Semi-Autonomous Neural ODEs and Applications In preparation (2024).

[8] Kang Liu, Ziqi Wang, and E. Zuazua. Game theory in federated learning: A potential game perspective. In preparation (2024).

[9] Kang Liu and Enrique Zuazua. On the sparse representation of Neural Networks In preparation (2024).

[10] Domènec Ruiz-Balet, Elisa Affili, and Enrique Zuazua. Interpolation and approximation via Momentum ResNets and Neural ODEs. *Systems & Control Letters*, 162:105182, 2022.

[11] Domènec Ruiz-Balet and Enrique Zuazua. Neural ODE Control for Classification, Approximation, and Transport. *SIAM Review*, 65(3):735–773, 2023. doi: 10.1137/21M1411433.

[12] Domènec Ruiz-Balet and Enrique Zuazua. Control of neural transport for normalising flows. *Journal de Mathématiques Pures et Appliquées*, 181:58-90, 2024.

[13] Yongcun Song, Ziqi Wang, and Enrique Zuazua. Approximate and weighted data reconstruction attack in federated learning. *arXiv preprint arXiv:2308.06822*, 2023.

[14] Yongcun Song, Ziqi Wang, and Enrique Zuazua. Fedadmm-insa: An inexact and self-adaptive admm for federated learning. *arXiv preprint arXiv:2402.13989*, 2024.

[15] Enrique Zuazua. Control and Machine Learning. *SIAM News*, 55(8), October 2022.

FRENCH-GERMAN-SPANISH CONFERENCE ON OPTIMIZATION
Gijón, June 18-21, 2024
(pp. 124-128)

Compatible TOSets with POSets: an application to additive manufacturing

Policarpo Abascal[1], **Fernando Fueyo** [2], **Jorge Jiménez** [3], **Antonio Palacio** [4], **Maria Luisa Serrano** [5]

1. abascal@uniovi.es Universidad de Oviedo, España
2. fueyofernando@uniovi.es Universidad de Oviedo, España
3. meana@uniovi.es Universidad de Oviedo, España
4. palacioantonio@uniovi.es Universidad de Oviedo, España
5. mlserrano@uniovi.es Universidad de Oviedo, España

Abstract

Additive Manufacturing (AM) has become a widely used technique in 3D printing, but it has proven to be a very costly process, even when optimizing parameters in existing models. Due to the characteristics of AM, and in order to optimize its process, a new approach is introduced to the problem: the discretization of each layer to be printed. This involves establishing an order relation based on the sequence in which the layers should be printed. The valid orders for the execution of the process, referred to as compatible with the order relation, will be characterized. Additionally, algorithms will be provided to obtain new compatible orders from others that were already compatible, and strategies will be presented to optimally and efficiently reorder non-compatible orders, converting them into compatible ones.

1. Introduction

The presentation collects part of the ideas we developed to solve a problem presented to us by a company for optimizing the 3D-printing of an object. This process falls within the context of Additive Manufacturing (AM) in which, each object is created from a set of layers. The use of printing layers allows for the creation of objects with a virtually unlimited variety of geometries, adaptable to any requirements of the final product. Paired with the advantage of printing any imaginable geometry, it appears the drawback of the slowness and cost of this production process. The technology, energy and human resources employed have a very high cost, so minimizing processing time naturally becomes a desired goal for all companies using this production method.

As usual, the problem consists of two well-differentiated parts: Modeling and Resolution. The talk starts by explaining some results that have been found in modeling, and it will finish with others related to optimization. To model the problem were used binary relations, that means equivalence relations but more, Order Relations.To solve the problem, that is, to minimize the processing time, were used Genetic Algorithms.

Each of the layers of the object contains a large set of points. This set of points is the unique piece of information required to process the object, that means that having control over this set, turns into having control over the production of the object. To get this, it was necessary to order and classify these points in some way. The order in which the information is provided to the device is crucial since the execution time depends strongly on this arrangement.

After performing a series of classifications on the set of points, using certain order and equivalence relations, were obtained a Partially Ordered Set (POSet) with a computationally acceptable number (10-120) of elements (pieces). Observing the diagram associated with the POSet from the perspective of Graph Theory, the problem consists of a particular version of the Traveling Salesman Problem (TSP). This version is due to the idiosyncrasies of the machines we are working with; we might refer to it as the Constrained Traveling Salesman Problem (CTSP). To solve it, they are used genetic algorithm techniques.

2. Initial Definitions and Properties

Definition 2.1 A binary relation R defined on a set S is a subset of $S \times S$. If $(a, b) \in R$ it is said to be a is R-related to b. R is said to be an **order relation** or a **partial order relation** on S if it is:

- reflexive: $(a, a) \in R \; \forall a \in S$
- antisymmetric: $\forall a, b \in S$ if $(a, b) \in R$ and $(b, a) \in R$ then $a = b$
- transitive: $\forall a, b, c \in S$ if $(a, b) \in R$ and $(b, c) \in R$ then $(a, c) \in R$

 ISBN 978-84-10135-30-7

A set S with a partial order relation is denoted by (S, R) and is known as **Partial Ordered Set** or **POSet**. If (S, R) is a POSet, then $a, b \in S$ are said to be **comparable** if aRb or bRa.
Let R be a binary relation defined on a set S, it is said to be a **total order relation** if it is an order relation and all the elements of S are comparable. If R is a total order (S, R) is said to be a **Totally Ordered Set**, or **TOSet**.

Definition 2.2 Given a set S with n elements, and a bijection from $I = \{1, 2, \dots, n\}$ to S

$$\begin{array}{ccc} I & \rightarrow & S \\ i & \rightarrow & a_i \end{array}$$

This establishes an indexing by means I of the elements of S. Then, S is said to be an **I-indexed set** or an **indexed set**.

Definition 2.3 Given σ a permutation of elements of I

$$\begin{array}{rccc} \sigma : I & \rightarrow & I \\ i & \rightarrow & \sigma(i) \end{array}$$

an **ordering** or **permutation** of elements of S can be generated as

$$\begin{array}{ccc} I & \rightarrow & S \\ i & \rightarrow & a_{\sigma(i)} \end{array}$$

We can represent the permutation σ by the images of the bijection that σ defines from I to itself as $(\sigma(1), \sigma(2), \dots, \sigma(n))$.

Definition 2.4 Let S be an indexed set with $Card(S) = n$ and R an order relation defined on S. We say that a matrix $M = (m_{ij})_{n \times n}$, is the **adjacency matrix** of (S, R) if it satifies:

$$m_{R,ij} = \begin{cases} 1 & \text{if } a_i R a_j \\ 0 & \text{otherwise} \end{cases}$$

Obviously, the adjacency matrix depends on the ordering in which the elements are taken. Thus, for each permutation $(\sigma(1), \sigma(2), \dots, \sigma(n))$ of elements of S, a matrix will be obtained, denoted by M_R^σ, and whose elements are:

$$m_{R,ij}^\sigma = \begin{cases} 1 & \text{if } a_{\sigma(i)} R a_{\sigma(j)} \\ 0 & \text{otherwise} \end{cases}$$

When there is no doubt about the order relation, the adjacency matrix for the permutation defined by σ can be denoted M^σ, and, for simplicity, we denote by M the adjacency matrix for the main permutation $(1, 2, 3, \dots, n)$. We denote by $MS(R)$ the set of the adjacency matrices that represent the relation R defined on the set S.

Proposition 2.5 *Let be an indexed set $S = \{a_1, a_2, \dots, a_n\}$, the total order relation R such that*

$$a_i R a_j \text{ if and only if } j \leq i$$

i.e., the elements ordered from highest to lowest index, and the adjacency matrix M^σ for R of a permutation $\sigma = (\sigma(1), \dots, \sigma(n))$ of elements of S, then for all $i_0 \in \{1, \dots, n\}$

$$\sum_{j=1}^{n} m_{j\, i_0}^\sigma = n - \sigma(i_0) + 1 \qquad \sum_{j=1}^{n} m_{i_0\, j}^\sigma = \sigma(i_0)$$

Proposition 2.6 *Given an indexed set S with n elements and an order relation R*

$$1 \leq Card(MS(R) \leq n!$$

If the order relation is total then $Card(MS(R)) = n!$

3. Compatibility

Definition 3.1 Given a permutation $\sigma = (\sigma(1), \sigma(2), \dots, \sigma(n))$ of elements of $I = \{1, 2, \dots, n\}$, we define the **relation induced by** σ **on** S and denote it by T_σ the relation defined as:

$$(a_i, a_j) \in T_\sigma \Leftrightarrow \sigma^{-1}(j) \leq \sigma^{-1}(i)$$

It is easy to see that, thus defined, this is a total order relation on S.

Definition 3.2 Let $S = \{a_1, a_2, \dots, a_n\}$ be an indexed set and let R be an order relation defined on S. A permutation σ of the elements of S is said to be **compatible with the relation** R if $R \subseteq T_\sigma$. We denote the set of permutations compatible with the relation R by $C(S, R)$.

Theorem 3.3 *Let (S, R) be an ordered indexed set and σ a permutation of elements of I; then*

$$\sigma \text{ is compatible with the relation } R \Leftrightarrow M_R^\sigma \text{ is lower triangular.}$$

Theorem 3.4 *Given an indexed POSet (S, R) with n elements, there is always a compatible permutation.*

Proof In a finite POSet, there always exist maximal elements. Let's assume there are μ_1 of these maximal elements.

Consider these maximal elements of (S, R),

$$M_{11}, M_{12}, \dots, M_{1\mu_1}$$

denoting by $M_{11} = a_{\sigma(1)}, M_{12} = a_{\sigma(2)}, \dots, M_{1\mu_1} = a_{\sigma(\mu_1)}$, we construct

$$(\sigma(1), \sigma(2), \dots, \sigma(\mu_1))$$

which verifies that if $i, j \in \{1, 2, \dots, \mu_1\}$, $a_{\sigma(i)}$ and $a_{\sigma(j)}$ are not comparable.

Let us now consider the set $S_1 = S - \{a_{\sigma(1)}, a_{\sigma(2)}, \dots, a_{\sigma(\mu_1)}\}$, and the restriction of R on S_1 that we denote R_1. As in the previous step, let's suppose that there are r_2 maximal elements, and, denoting $\mu_2 = \mu_1 + r_2$ and

$$M_{21} = a_{\sigma(\mu_1+1)}, M_{22} = a_{\sigma(\mu_1+2)}, \dots, M_{2r_2} = a_{\sigma(\mu_2)},$$

we add them to the previously constructed permutation, obtaining

$$(\sigma(1), \sigma(2), \dots, \sigma(\mu_1), \sigma(\mu_1 + 1), \dots, \sigma(\mu_2))$$

that verifies

- If $1 \leq i, j \leq \mu_1 \Rightarrow a_{\sigma(i)}$ and $a_{\sigma(j)}$ are not comparable
- If $\mu_1 < i, j \leq \mu_2 \Rightarrow a_{\sigma(i)}$ and $a_{\sigma(j)}$ are not comparable
- If $1 \leq i \leq \mu_1 < j \leq \mu_2 \Rightarrow$ as $a_{\sigma(j)}$ is maximal in S_1, $a_{\sigma(i)}$ is maximal in S and $S_1 \subseteq S$ therefore $a_{\sigma(i)}$ and $a_{\sigma(j)}$ are not comparable or $a_{\sigma(j)} R a_{\sigma(i)}$.

Repeating the process $k - 1$ times considering the set $S_k = S - \{a_{\sigma(1)}, \dots, a_{\sigma(\mu_k)}\}$ and taking the maximal elements of the poset (S_k, R_k) being R_k, the restriction of R to the set S_k, we will obtain, after a finite number of steps, a permutation of the n elements of S

$$\sigma = (\sigma(1), \sigma(2), .., \sigma(\mu_1), \sigma(\mu_1 + 1), .., \sigma(\mu_2), .., \sigma(\mu_k), \sigma(\mu_k + 1), .., \sigma(n))$$

which is compatible with the relation R by construction. □

Definition 3.5 Let be an indexed ordered set (S, R) with n elements, σ_1 and σ_2, permutations of elements of S and $k \in \{1, \dots, n-1\}$, we call the k-cut offspring permutation of σ_1 and σ_2 the permutation γ defined as:

$$\gamma = (\sigma_1(1), \sigma_1(2), \dots, \sigma_1(k), \sigma_2(i_1), \dots, \sigma_2(i_{n-k}))$$

where for all $h \in \{i_1, i_2, \dots, i_{n-k}\}$ such that $i_1 < i_2 < \dots < i_{n-k}$ then $\sigma_2(h) \notin \{\sigma_1(1), \dots, \sigma_1(k)\}$.

Theorem 3.6 *Given an indexed ordered set (S, R) with n elements and $k \in \{1, \dots, n-1\}$, the k-cut offspring permutation of two permutations, σ_1 and σ_2, compatible R is a permutation compatible with R.*

Proof Let be $\sigma_1 = (\sigma_1(1), \dots, \sigma_1(n)), \sigma_2 = (\sigma_2(1), \dots, \sigma_2(n))$ two compatible permutation and $k \in \{1, \dots, n-1\}$.

The k-cut offspring permutation is

$$\gamma = (\sigma_1(1), \sigma_1(2), \dots, \sigma_1(k), \sigma_2(i_1), \dots, \sigma_2(i_{n-k}))$$

where for all $h \in \{i_1, i_2, \dots, i_{n-k}\}$ such that $i_1 < i_2 < \dots < i_{n-k}$ then $\sigma_2(h) \notin \{\sigma_1(1), \dots, \sigma_1(k)\}$.

Let's denote $S^1 = \{a_{\sigma_1(1)}, \dots, a_{\sigma_1(k)}\}$ and $S^2 = \{a_{\sigma_2(1)}, \dots, a_{\sigma_2(k)}\}$

- If $S^1 = S^2$, the elements belonging to S^1 are compatible with each other in the resulting permutation due to their presence in the compatible permutation σ_1, and the remaining elements $S - S^1$ with each other as well, because they are in σ_2.

  The elements belonging to S^1 are also compatible with those in $S - S^1$ by verifying the compatibility of σ_2.

  So, in this case we have a resulting permutation compatible with the relation.

- If $S^1 \neq S^2$
  - the elements of S^1 and those of $S - S^1$, due to the compatibility of σ_1 and σ_2, respectively, are compatible with each other in the resulting permutation;
  - if $a \in S^1$ and $b \in (S - S^1)$,

    $$a = a_{\sigma_1(j_a)} = a_{\sigma_2(i_a)} \text{ and } j_a < k \quad b = a_{\sigma_1(j_b)} = a_{\sigma_2(i_b)} \text{ and } j_b > k$$

    * if $b \notin S^2 \longrightarrow i_b > k$ then a and b are compatible in the resulting permutation;
    * if $b \in S^2 \longrightarrow i_b < k$
      · if $i_a < i_b$, they are in the same order in both permutations and are therefore compatible in the resulting permutation.
      · if $i_b < i_a$, as $j_a < k < j_b$, then they are interchanged in both permutations and therefore, by Proposition **??**, they are not comparable and, therefore, are compatible in the resulting permutation.

  So, in this case, we also have a resulting permutation compatible with the order relation.

Then we can conclude that the permutation resulting from two compatible permutations with the relation R is also a compatible one. □

Definition 3.7 The procedure described in Definition 3.5 can be extended recursively to the case of $m > 2$ permutations and a partition, $k = (k_1, \dots, k_m)$, of n, that is $\forall i \in \{1, \dots, m\} k_i \in \{1, \dots, n-1\}$ and $\sum_{i=1}^{m} k_i = n$. Given $\sigma_i = (\sigma_i(1), \sigma_i(2), \dots, \sigma_i(n))$, $i \in \{1, \dots, m\}$ permutations of elements of S and $k = (k_1, \dots, k_m = n - \sum_{i=1}^{m-1} k_i)$, we construct γ_m as follows

$$\begin{cases} \gamma_2 = k_1\text{-cut offspring permutation of } \sigma_1 \text{ and } \sigma_2 \\ \gamma_i = \left(\sum_{i=1}^{i-1} k_i\right)\text{-cut offspring permutation of } \gamma_{i-1} \text{ and } \sigma_i, \quad \text{if } i \in \{3, \dots m\} \end{cases}$$

and we call it $(k_1, k_2, \dots, k_{m-1})$-cut offspring permutation of $\sigma_1, \sigma_2, \dots, \sigma_m$. γ_m is that which the elements of the positions between $\sum_{j=1}^{i-1} k_j$ and $\sum_{j=1}^{i} k_j$ are the first k_i elements of permutation σ_i that are not in $\bigcup_{j=1}^{i-1} \{\sigma_j(i_{j_1}), \dots, \sigma_j(i_{j_{k_j}})\}$.

Theorem 3.8 *Given an indexed ordered set (S, R) with n elements, the resulting permutation of m permutations compatible with R, $k = (k_1, \dots, k_m)$ a partition of n, that is, as in the Definition 3.7, $\sum_{i=1}^{m} k_i = n$, is a permutation compatible with the relation R.*

Acknowledgements

This work has been partially supported by the collaborative project *FUO*-115-22.

References

[1] Birkhoff G.: Lattice Theory, American Mathematical Society, Providence. (1948)

[2] Caspard N., Leclerc B., Monjardet B.: Finite Ordered Sets. Concepts, Results and Uses, Encyclopedia of Mathematics and Its Application, 144. Cambridge University Press, Cambridge (2012)

[3] Chowdhury S., Yadaiah N., Prakash C., Ramakrishna S., Dixit S., Gupta L R., Buddhi D.: Laser powder bed fusion: a state-of-the-art review of the technology, materials, properties & defects, and numerical modelling, J. Market. Res. **20**, pp. 2109-2172. (2022) https://doi.org/10.1016/j.jmrt.2022.07.121

[4] Garg V. K.: Introduction to Lattice Theory with Computer Science Applications. Wiley, Hoboken (2015)

[5] Khorasani A., Gibson I., Veetil J. K. et al.: A review of technological improvements in laser-based powder bed fusion of metal printers. Int. J. Adv. Manuf. Technol. **108**:191-209, (2020) https://doi.org/10.1007/s00170-020-05361-3

[6] Olleak A., Xi Z.: Efficient LPBF process simulation using finite element modeling with adaptive remeshing for distortions and residual stresses prediction, Manuf. Lett. **24**, 140-144 (2020). https://doi.org/10.1016/j.mfglet.2020.05.002. (ISSN 2213-8463)

[7] Viguerie A., Bertoluzzo S., Auricchio F.: A Fat boundary-type method for localized nonhomogeneous material problems. Comput. Methods Appl. Mech. Eng. (2020). https://doi.org/10.1016/j.cma.2020.112983

Author Index

T

W

Y

Z