DE GRUYTER

# OPTIMIZATION AND CONTROL FOR PARTIAL DIFFERENTIAL EQUATIONS

## UNCERTAINTY QUANTIFICATION, OPEN AND CLOSED-LOOP CONTROL, AND SHAPE OPTIMIZATION

*Edited by Roland Herzog, Matthias Heinkenschloss, Dante Kalise, Georg Stadler, Emmanuel Trélat*

RICAM
JOHANN-RADON-INSTITUTE FOR COMPUTATIONAL AND APPLIED MATHEMATICS

ÖAW
AUSTRIAN ACADEMY OF SCIENCES

RADON SERIES ON COMPUTATIONAL AND APPLIED MATHEMATICS 29

Copyright 2022. De Gruyter. All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

EBSCO Publishing : eBook Collection (EBSCOhost) - printed on 2/10/2023 3:19 PM via AN: 3[illegible]06 ; Roland Herzog, Matthias Heinkenschloss, Dante Kalise, Georg Stadler, Emman[illegible]Trlat.; Optimization and Control for Partial Differential Equations : Uncer[illegible]ty Quantification, Open and Closed-loop Control, and Shape Optimization
Accou[illegible]ns335141

Roland Herzog, Matthias Heinkenschloss, Dante Kalise, Georg Stadler, and Emmanuel Trélat (Eds.)

**Optimization and Control for Partial Differential Equations**

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

# Radon Series on Computational and Applied Mathematics

Managing Editor
Ulrich Langer, Linz, Austria

Editorial Board
Hansjörg Albrecher, Lausanne, Switzerland
Ronald H. W. Hoppe, Houston, Texas, USA
Karl Kunisch, Linz/Graz, Austria
Harald Niederreiter, Linz, Austria
Otmar Scherzer, Linz/Vienna, Austria
Christian Schmeiser, Vienna, Austria

## Volume 29

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

# Optimization and Control for Partial Differential Equations

Uncertainty quantification, open and closed-loop control, and shape optimization

Edited by
Roland Herzog, Matthias Heinkenschloss, Dante Kalise, Georg Stadler, and Emmanuel Trélat

DE GRUYTER

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Editors**
Prof. Dr. Roland Herzog
Heidelberg University
Interdisciplinary Center for Scientific Computing
Im Neuenheimer Feld 205
69120 Heidelberg
Germany
roland.herzog@iwr.uni-heidelberg.de

Prof. Dr. Georg Stadler
New York University
Courant Inst. of Mathematical Sciences Library
251 Mercer Street
New York, NY 10012-1185
USA
stadler@cims.nyu.edu

Prof. Dr. Matthias Heinkenschloss
Rice University
Department of Computational and Applied Mathematics
Houston, Main 77005-1827
USA
heinken@rice.edu

Prof. Dr. Emmanuel Trélat
Laboratoire J.-L. Lions
Sorbonne-Université – BC187
4 pl. Jussieu
75252 Paris Cedex 05
France
emmanuel.trelat@sorbonne-universite.fr

Prof. Dante Kalise
Imperial College London
Dept. of Mathematics
South Kensington Campus
London SW7 2AZ
UK
d.kalise-balza@imperial.ac.uk

ISBN 978-3-11-069596-0
e-ISBN (PDF) 978-3-11-069598-4
e-ISBN (EPUB) 978-3-11-069600-4
ISSN 1865-3707

**Library of Congress Control Number: 2021950264**

**Bibliographic information published by the Deutsche Nationalbibliothek**
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

© 2022 Walter de Gruyter GmbH, Berlin/Boston
Typesetting: VTeX UAB, Lithuania
Printing and binding: CPI books GmbH, Leck

www.degruyter.com

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

# Preface

*The optimist proclaims that we live in the best of all possible worlds;*
*and the pessimist fears this is true.*

James Branch Cabell, The Silver Stallion

Over the last century, our information society has witnessed the emergence of a new method of scientific inquiry underpinned by Applied Mathematics. Our understanding of complex systems such as the Earth's weather, the financial market, or social networks, to name a few, is based around mathematical modeling, simulation, and optimization. Modeling and simulation deal with an accurate identification and replication of the underlying processes. Optimization addresses the characterization and computation of actions to achieve a best possible outcome.

Mathematical optimization has been present throughout the history of humankind. From Queen Dido's legend and the isoperimetric problem, going through Fermat's principle of least time in 1662, until the 20th century development of optimal control theory by Bellman and Pontryagin, the synthesis of optimal actions has been a formidable mathematical challenge. The present volume is a nonexhaustive, albeit sufficiently broad, timely account of some of the most prominent current research directions in optimization and optimal control.

This volume features contributions presented at the Special Semester on Optimization, organized by Karl Kunisch and Ekkehard Sachs, held from 14th October until 11th December 2019 at the Johann Radon Institute for Computational and Applied Mathematics (RICAM) in Linz, Austria. Over 140 speakers from all parts of the globe presented their latest research, showcased around five thematic workshops:

**New Trends in PDE-Constrained Optimization** (organizers: Roland Herzog and Emmanuel Trélat)

This workshop focused on recent advances in the wide and growing field of PDE-constrained optimization. Participants discussed recent developments in modeling, theory, and numerical techniques. Challenging applications ranged over a broad variety of fields, including additive manufacturing, biology, geophysics, fluid flow, medical imaging, solid mechanics, and natural hazards.

**Optimal Control and Optimization for Nonlocal Models** (organizers: Max Gunzburger and Marta D'Elia)

Participants at this workshop discussed a broad range of models, analytical and numerical techniques, and applications involving nonlocal models, with an emphasis on control and optimization. Examples included fractional differential equations, learning problems for nonlocal models, and nonlocal inverse problems.

**Optimization and Inversion Under Uncertainty** (organizers: Matthias Heinkenschloss and Georg Stadler)

This workshop showcased theoretical and algorithmic advances in the areas of optimization under uncertainty and Bayesian inversion governed by complex physi-

https://doi.org/10.1515/9783110695984-201

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

cal systems. Problems in these areas arise in many science and engineering applications where we need to make decisions, compute designs, or specify controls under uncertainty, or where we need to infer parameters such as spatially distributed conductivities or permeabilities from measurements. Participants in this workshop presented advances in the development and integration of mathematical and computational tools from uncertainty quantification, optimization, control, and PDEs to solve these challenging problems.

**Nonsmooth Optimization** (organizer: Christian Clason)

This workshop focused on generalized calculus, optimality conditions, and algorithms for various classes of nonsmooth optimization problems. A variety of examples were discussed, including Nash equilibria, bilevel and multiobjective optimization, and quasi-variational inequalities.

**Feedback Control** (organizers: Dante Kalise and Karl Kunisch)

The focus of this workshop comprised theoretical and numerical aspects of feedback control. Topics discussed included feedback control and controllability of systems governed by PDEs, optimal feedback control, agent-based models, and computational methods for high-dimensional feedback synthesis.

**Conic and Copositive Optimization** (organizer: Mirjam Dür)

Conic optimization studies optimization problems where the decision variable is required to lie in some closed convex cone. Notable examples include the nonnegative orthant, the second-order cone, the cone of positive semidefinite matrices, and the cone of copositive matrices. At the workshop, participants presented analytical results and algorithmic advances.

We take this opportunity to thank Karl Kunisch, Ekkehard Sachs, and the staff at RICAM, notably Annette Weihs, for their support in the organization of the special semester on optimization. Financial support by RICAM and the Austrian Academy of Sciences is gratefully acknowledged. We also wish to thank Nadja Schedensack and the editorial staff at De Gruyter for their support in the production of this volume.

Houston, Heidelberg, London, New York, Paris, 2021

Matthias Heinkenschloss
Roland Herzog
Dante Kalise
Georg Stadler
Emmanuel Trélat

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

# Contents

**Preface —— V**

Marco Bernreuther, Georg Müller, and Stefan Volkwein
**1 Reduced basis model order reduction in optimal control of a nonsmooth semilinear elliptic PDE —— 1**

Arthur Bottois
**2 Pointwise moving control for the 1-D wave equation —— 33**

Martin Gugat and Michael Herty
**3 Limits of stabilizability for a semilinear model for gas pipeline flow —— 59**

Luis Almeida, Jesús Bellver Arnau, Michel Duprez, and Yannick Privat
**4 Minimal cost-time strategies for mosquito population replacement —— 73**

Luis Almeida, Jorge Estrada, and Nicolas Vauchelet
**5 The sterile insect technique used as a barrier control against reinfestation —— 91**

Evelyn Herberg and Michael Hinze
**6 Variational discretization approach applied to an optimal control problem with bounded measure controls —— 113**

Adrian Hirn and Winnifried Wollner
**7 An optimal control problem for equations with *p*-structure and its finite element discretization —— 137**

Ulrich Langer, Olaf Steinbach, Fredi Tröltzsch, and Huidong Yang
**8 Unstructured space-time finite element methods for optimal sparse control of parabolic equations —— 167**

Carolin Dirks and Benedikt Wirth
**9 An adaptive finite element approach for lifted branched transport problems —— 189**

Florian Feppon
**10 High-order homogenization of the Poisson equation in a perforated periodic domain —— 237**

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Jérôme Lemoine and Arnaud Münch
**11 Least-squares approaches for the 2D Navier–Stokes system —— 285**

Gontran Lance, Emmanuel Trélat, and Enrique Zuazua
**12 Numerical issues and turnpike phenomenon in optimal shape design —— 343**

Gabriela Marinoschi
**13 Feedback stabilization of Cahn–Hilliard phase-field systems —— 367**

Philipp A. Guth, Claudia Schillings, and Simon Weissmann
**14 Ensemble Kalman filter for neural network-based one-shot inversion —— 393**

Joost A. A. Opschoor, Christoph Schwab, and Jakob Zech
**15 Deep learning in high dimension: ReLU neural network expression for Bayesian PDE inversion —— 419**

**Index —— 463**

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Marco Bernreuther, Georg Müller, and Stefan Volkwein

# 1 Reduced basis model order reduction in optimal control of a nonsmooth semilinear elliptic PDE

**Abstract:** In this paper, we consider an optimization problem governed by a nonsmooth semilinear elliptic partial differential equation. We apply a reduced order approach to obtain a computationally fast and certified numerical solution approach. Using the reduced basis method and efficient a posteriori error estimation for the primal and dual equations, we develop an adaptive algorithm and successfully test it for several numerical examples.

**Keywords:** nonsmooth optimization, nonsmooth semilinear elliptic equations, semismooth Newton, reduced basis method, error estimation

**MSC 2010:** 35J20, 49K20, 49M05, 65K10

## 1.1 Introduction

In this paper, we consider the optimal control problem governed by a nonsmooth semilinear elliptic partial differential equation (PDE)

$$\begin{aligned} &\min_{(y,\mu)} \mathcal{J}(y,\mu) = j(y) + \frac{\sigma}{2}\|\mu\|_{\mathrm{A}}^2 \\ &\quad \text{s. t. } (y,\mu) \in V \times \mathbb{R}^p \text{ satisfies } -\Delta y + \max\{0,y\} = \mathcal{B}\mu \text{ in } V', \end{aligned} \tag{\mathbf{P}}$$

where $\mu$ denotes a parameter that acts as a control on the right-hand side, and $y$ is the state. We endow $V := H_0^1(\Omega)$ with the usual inner product

$$\langle \varphi, \phi \rangle_V := \int_\Omega \nabla\varphi \cdot \nabla\phi + \varphi\phi \, \mathrm{d}\boldsymbol{x} \quad \text{for } \varphi, \phi \in V$$

**Acknowledgement:** This research was supported by the German Research Foundation (DFG) under grant number VO 1658/5-2 within the priority program "Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization" (SPP 1962).

**Marco Bernreuther, Georg Müller, Stefan Volkwein,** University of Konstanz, Department of Mathematics and Statistics, WG Numerical Optimization, Universitätsstraße 10, 78457 Konstanz, Germany, e-mails: marco.bernreuther@uni-konstanz.de, georg.mueller@uni-konstanz.de, stefan.volkwein@uni-konstanz.de

https://doi.org/10.1515/9783110695984-001

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

and the induced norm $\|\cdot\|_V := \langle\cdot,\cdot\rangle_V^{1/2}$. Its topological dual space is denoted $V'$. Our assumptions on the data throughout the paper are as follows.

**Assumption 1.1.1.**

1) $\Omega \subset \mathbb{R}^d$, $d \geq 1$, is a bounded domain that is either convex or possesses a $C^{1,1}$-boundary (cf. [10, Section 6.2]),
2) $j\colon V \to \mathbb{R}$ is weakly lower semicontinuous, twice continuously differentiable, and bounded from below,
3) $\sigma > 0$,
4) $p \in \mathbb{N} \setminus \{0\}$, $\mathsf{A} \in \mathbb{R}^{p\times p}$ is diagonal positive definite, and $\|\cdot\|_{\mathsf{A}} := \langle \mathsf{A}\cdot,\cdot\rangle^{1/2}$,
5) $\mathcal{B}\colon \mathbb{R}^p \to L^2(\Omega)$ is linear and bounded, the pairwise intersection of the sets $\{b_i \neq 0\}$ of $b_i := \mathcal{B}(e_i) \in L^2(\Omega)$, where $e_i$, $i = 1,\dots,p$, denote the unit vectors in $\mathbb{R}^p$, are Lebesgue nullsets, and all $b_i$ are nonzero.

The governing constraint is a rather well-understood semilinear elliptic PDE; see, e. g., [8, 9]. It features a "level-1-type" nonsmooth Nemytski operator that induces a nonsmoothness in the solution operator, which remains Lipschitz continuous and Hadamard differentiable. Depending on the specific result, our examinations can, of course, be generalized for additional parameters or different Nemytski operators. We refer to Assumption 1.1.1 and Section 1.2 for the exact setting chosen in this paper. Some examples of problems related to the nonsmooth PDE arise, for instance, in mechanics, plasma physics, and the context of certain combustion processes; see, e. g., [16, 21, 22, 25] for possible areas of application.

The goal of the present paper is the development, analysis, and numerical realization of efficient, fast, and certified solution algorithms for a first-order system of (**P**). Of course, for the implementation, a discretization of the state space $V$ and the PDE constraint is required. Utilizing a standard finite element (FE) method for the discretization of the first-order system of (**P**), we derive a complex nonlinear and nonsmooth large-scale system. To obtain fast numerical solution methods, we apply a reduced order approach employing the reduced basis (RB) method; see, e. g., [3, 13, 18, 20].

For the construction of an accurate and efficient RB scheme, a posteriori error analysis is required. Recall that the RB method is especially efficient for parameterized linear-quadratic optimal control problems; see, e. g., [15, 17]. For RB results in the framework of nonlinear elliptic PDEs, we refer, e. g., to [5, 11, 23] and to the recent work [14].

Results concerning model order reduction (MOR) for nonsmooth PDE constrained optimization are rarely found in the literature, with the only contribution, to the best of the authors' knowledge, being [4], where an offline/online based (greedy) reduced basis framework for the PDE constraint in (**P**) is investigated. The results showed that the RB method, combined with empirical interpolation techniques (see, e. g., [2, 6, 7]), can improve the efficiency of solving PDEs, however, with nonsmooth effects being

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

the limiting factor for the quality of the reduced-order approximations. Since the constraint is part of the first-order system for (**P**), we extend the results obtained in the thesis [4] and present an RB approach for the constraint justified by efficient a posteriori error analysis. This way, we obtain a significant reduction of the CPU times compared to a standard FE discretization. For the optimization problem (**P**), we consider first-order conditions based on those derived in [9], i. e., the system

$$-\Delta\bar{y} + \max\{0, \bar{y}\} = \mathcal{B}\bar{u}, \tag{1.1a}$$

$$-\Delta\bar{p} + \mathbb{1}_{\{\bar{y}>0\}}\bar{p} = j'(\bar{y}), \tag{1.1b}$$

$$\mathcal{B}^*\bar{p} + \sigma\mathsf{A}\mu = 0, \tag{1.1c}$$

and the corresponding pseudo-semismooth Newton (PSN) scheme. In (1.1b) the symbol $\mathbb{1}$ denotes the indicator function, and in (1.1c) the operator $\mathcal{B}^* \in \mathcal{L}(L^2(\Omega), \mathbb{R}^p)$ is the adjoint of $\mathcal{B}$. Starting from a classical offline/online RB concept, we introduce a novel adaptive combined RB-PSN scheme, which does not require an offline phase and improves the RB approximation quality simultaneously with the solution iterations using local information along the iterates.

The paper is organized as follows. In Section 1.2, we introduce and analyze the RB method for the parameterized nonsmooth PDE constraint of the optimization problem, including residual-based error estimation. Numerical tests illustrate the drastic reduction of the CPU times compared to a piecewise linear FE scheme. Section 1.3 covers the offline/online and the adaptive RB approach for system (1.1). We explain the adaptive RB approach in detail, derive a residual-based error indicator for the first-order system, and illustrate the efficiency and accuracy of the method by several numerical experiments. Finally, we draw some conclusions in Section 1.4.

## 1.2 MOR for the state equation

The aim of this section is to establish classical MOR results based on the RB approach, such as error estimators, a priori estimates, and convergence theory, for the nonsmooth PDE constraint. The crucial ingredients for this will be the monotonicity of the max-operator and the Lipschitz continuity of the solution operator to the PDE. The presented results are mainly based on [4].

Let us first fix the exact analytical framework for the FE and RB analysis of the PDE and recall some known results. We consider the more general parameterized boundary value problem

$$c(\mu)\langle\nabla y, \nabla\varphi\rangle_{L^2} + a(\mu)\langle\max\{0, y\}, \varphi\rangle_{L^2} = \langle f(\mu), \varphi\rangle_{V',V} \quad \forall\varphi \in W, \tag{1.2}$$

where $\mu$ is in a parameter set $\mathcal{P} \subset \mathbb{R}^p$, $p \in \mathbb{N} \setminus \{0\}$, and $W$ denotes a closed subspace of $V$; in the FE setting, this is the finite element space, and in the MOR framework, this

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

will be a low-dimensional subspace spanned by the RB functions. For the remainder of this section, we additionally assume the following:

**Assumption 1.2.1.**
1) $\mathcal{P} \subset \mathbb{R}^p$, $p \in \mathbb{N} \setminus \{0\}$, is nonempty and closed,
2) $c\colon \mathcal{P} \to \mathbb{R}$ is $L_c$-Lipschitz continuous, positive, and uniformly bounded away from zero,
3) $a\colon \mathcal{P} \to \mathbb{R}$ is $L_a$-Lipschitz continuous and nonnegative,
4) $f\colon \mathcal{P} \to V'$ is $L_f$-Lipschitz continuous,
5) $\mathcal{P}$ is compact, or both $a$ and $c$ are constant.

The following proposition sums up the standard existence and uniqueness result for solutions to equation (1.2) given the assumptions made above and provides some additional information on the Lipschitz constant.

**Proposition 1.2.2.** *Let Assumptions* 1.1.1 *and* 1.2.1 *be satisfied. For every $\mu \in \mathcal{P}$, there exists a unique solution $y \in W$ to equation* (1.2). *The induced solution operator $\mathcal{S}_W\colon \mathcal{P} \to W$ is $L_{\mathcal{S}}$-Lipschitz continuous with constant*

$$L_{\mathcal{S}} = \frac{C_1}{C_2}\left(\frac{L_c + L_a C_1}{C_2} \sup_{\mu\in\mathcal{P}} \|f(\mu)\|_{V'} + L_f\right) \tag{1.3}$$

*independently of $W$, where $C_P > 0$ denotes the Poincaré constant, $C_1 = 1 + C_P^2$, and $C_2 = \inf\{c(\mu) \mid \mu \in \mathcal{P}\}$.*

*Proof.* Though the result is fairly standard, we will give a short outline of the proof for completeness. The existence and uniqueness is guaranteed by the Browder–Minty theorem. For the proof of the Lipschitz continuity, we set $y_1 := y(\mu_1)$ and $y_2 := y(\mu_2)$. Without loss of generality, we assume that $\|y_1 - y_2\|_V > 0$; otherwise, the statement is trivial. We subtract the PDEs

$$c(\mu_1)\langle\nabla y_1, \nabla\varphi\rangle_{L^2} + a(\mu_1)\langle\max\{0, y_1\}, \varphi\rangle_{L^2} = \langle f(\mu_1), \varphi\rangle_{V',V} \quad \forall\varphi \in W, \tag{1.4}$$

$$c(\mu_2)\langle\nabla y_2, \nabla\varphi\rangle_{L^2} + a(\mu_2)\langle\max\{0, y_2\}, \varphi\rangle_{L^2} = \langle f(\mu_2), \varphi\rangle_{V',V} \quad \forall\varphi \in W, \tag{1.5}$$

add a zero, and rearrange the terms to obtain

$$\begin{aligned}&\langle c(\mu_1)\nabla y_1 - c(\mu_2)\nabla y_2, \nabla\varphi\rangle_{L^2} + a(\mu_1)\langle\max\{0, y_1\} - \max\{0, y_2\}, \varphi\rangle_{L^2}\\ &\quad + \big(a(\mu_1) - a(\mu_2)\big)\langle\max\{0, y_2\}, \varphi\rangle_{L^2} = \langle f(\mu_1) - f(\mu_2), \varphi\rangle_{V',V}.\end{aligned}$$

Now we test with $\varphi = y_1 - y_2 \in W$, use the monotonicity of the max operator, the Cauchy–Schwarz inequality, and nonnegativity of $a$ to obtain that

$$\begin{aligned}c(\mu_1)\|\nabla(y_1 - y_2)\|_{L^2}^2 \le{}& |a(\mu_1) - a(\mu_2)|\|\max\{0, y_2\}\|_{L^2}\|y_1 - y_2\|_{L^2}\\ &+ \|f(\mu_1) - f(\mu_2)\|_{V'}\|y_1 - y_2\|_V + |c(\mu_1) - c(\mu_2)|\|\nabla y_2\|_{L^2}\|\nabla(y_1 - y_2)\|_{L^2}.\end{aligned}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

The Poincaré inequality yields the $V$-estimate

$$\frac{c(\mu_1)}{C_1}\|y_1 - y_2\|_V^2 \leq |a(\mu_1) - a(\mu_2)|\|\max\{0, y_2\}\|_{L^2}\|y_1 - y_2\|_{L^2} + \|f(\mu_1) - f(\mu_2)\|_{V'}\|y_1 - y_2\|_V + |c(\mu_1) - c(\mu_2)|\|\nabla y_2\|_{L^2}\|\nabla(y_1 - y_2)\|_{L^2},$$

where the Lipschitz continuity of the coefficient functions gives

$$\begin{aligned}&\frac{c(\mu_1)}{C_1}\|y_1 - y_2\|_V^2 \\ &\quad \leq (L_a\|y_2\|_{L^2} + L_f + L_c\|\nabla y_2\|_{L^2})\|y_1 - y_2\|_V\|\mu_1 - \mu_2\|_{\mathbb{R}^p}.\end{aligned} \tag{1.6}$$

Applying the Poincaré inequality once more, we obtain the estimate

$$\|y_2\|_V^2 \leq C_1\|\nabla y_2\|_{L^2}^2. \tag{1.7}$$

We test (1.5) with $\varphi = y_2$ and end up with

$$\langle c(\mu_2)\nabla y_2, \nabla y_2\rangle_{L^2} + \underbrace{\langle a(\mu_2)\max\{0, y_2\}, y_2\rangle_{L^2}}_{\geq 0} = \langle f(\mu_2), y_2\rangle_{V',V}$$

due to the monotonicity of the max operator, and thus

$$c(\mu_2)\|\nabla y_2\|_{L^2}^2 \leq \langle f(\mu_2), y_2\rangle_{V',V}. \tag{1.8}$$

Now combining equations (1.7) and (1.8), we obtain that

$$\|y_2\|_V \leq \frac{C_1}{c(\mu_2)}\|f(\mu_2)\|_{V'}.$$

We combine that with (1.6) and (1.8), divide by $\|y_1 - y_2\|_V$, use the positivity of the lower bound of $c$, and rearrange terms to obtain that

$$\|y_1 - y_2\|_V \leq \frac{C_1}{C_2}\left(\frac{L_c + L_aC_1}{C_2}\|f(\mu_2)\|_{V'} + L_f\right)\|\mu_1 - \mu_2\|_{\mathbb{R}^p}.$$

If $a$ and $c$ are constant, then $L_a = L_c = 0$, which yields (1.3) with the usual convention of $0 \cdot \infty = 0$. Otherwise, $\mathcal{P}$ is compact, and we can form the supremum, which is attained as a maximum due to the compactness. □

**Remark 1.2.3.**

1) The Lipschitz constant $L_{\mathcal{S}}$ is explicitly computable for a given domain $\Omega$ whose Poincaré constant $C_P$ is known. This can be used to get the quantities in the convergence rate of the RB-to-FE distance; cf. Theorem 1.2.7.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

2) Higher regularity of the right-hand side $f$ induces higher regularity of the solutions. In this section, we are not concerned with these effects. The upcoming optimization section, however, is quite reliant on the corresponding effect in the optimization formulation.
3) Note that the solution operator $\mathcal{S}_W$ is generally not Gâteaux differentiable (cf. [9, Section 6]), but Hadamard differentiable for constant $a$ and $c$, as shown in [8, Theorem 22].

### 1.2.1 RB analysis for the state equation

For the RB analysis, we will assume that we are able to approximate the analytical solution of (1.2) with $W = V$ arbitrarily well using finite-dimensional FE spaces $V_h \subset V$ endowed by the $V$-topology. Accordingly, the solution $y_h(\mu)$ of (1.2) with $W = V_h$ will be considered the reference solution. We follow the classical offline/online splitting approach (see, e. g., [12, 18]), where the RB space is generated via a greedy algorithm in an offline phase so that the RB solutions with respect to this basis can be computed quickly later on in the online phase. The offline basis computation for an error indicator $\Delta(V_\ell, \mu)$ that satisfies

$$\|y_h(\mu) - y_\ell(\mu)\|_V \leq \Delta(V_\ell, \mu) \quad \text{for all } \mu \in \mathcal{P}_{\text{train}},$$

where $y_h(\mu) \in V_h$ and $y_\ell(\mu) \in V_\ell$ denote the FE and RB solutions, respectively, is described in Algorithm 1.1.

**Algorithm 1.1:** Greedy RB method for PDE.

**Require:** Discrete training set of parameters $\mathcal{P}_{\text{train}} \subset \mathcal{P}$, error tolerance $\varepsilon_{\text{tol}} > 0$
**Return :** RB parameters $\mathcal{P}_\ell$, reduced basis $\Psi_\ell$, RB space $V_\ell$
Set $\ell = 0$, $\mathcal{P}_0 = \emptyset$, $\Psi_0 = \emptyset$, $V_0 = \{0\}$;
**while** $\varepsilon_\ell := \max\{\Delta(V_\ell, \mu) \mid \mu \in \mathcal{P}_{\text{train}}\} > \varepsilon_{\text{tol}}$ **do**
  Compute $\mu_{\ell+1} \in \arg\max\{\Delta(V_\ell, \mu) \mid \mu \in \mathcal{P}_{\text{train}}\}$;
  Set $\mathcal{P}_{\ell+1} = \mathcal{P}_\ell \cup \{\mu_{\ell+1}\}$ and $\psi_{\ell+1} = y_h(\mu_{\ell+1})$;
  Orthonormalize $\psi_{\ell+1}$ against $\Psi_\ell$;
  Set $\Psi_{\ell+1} = \Psi_\ell \cup \{\psi_{\ell+1}\}$;
  Define $V_{\ell+1} = V_\ell \oplus \operatorname{span}(\psi_{\ell+1})$ and $\ell = \ell + 1$;

Note that the indicator $\Delta(V_\ell, \mu)$ is an essential component of the algorithm and should be easily evaluable. In our case the monotonicity of the max term allows us to derive a residual-based error estimator that only depends on the RB solution and can thus be evaluated without solving the possibly costly FE discretized PDE.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Proposition 1.2.4.** *Let Assumptions* 1.1.1 *and* 1.2.1 *hold, and let* $e_y(\mu) := y_h(\mu) - y_\ell(\mu)$. *Then*

$$\|e_y(\mu)\|_V \le \Delta_w^y(V_\ell, \mu) := \frac{C_1}{C_2} \|\mathrm{Res}_\ell^y(\mu)\|_{V_h'} \tag{1.9}$$

*for every* $\mu \in \mathcal{P}$ *with* $C_1 = 1 + C_P^2$, $C_2 = \inf\{c(\mu) \mid \mu \in \mathcal{P}\}$, *and the residual* $\mathrm{Res}_\ell^y \colon \mathcal{P} \to V_h'$ *given by*

$$\begin{aligned}\langle \mathrm{Res}_\ell^y(\mu), \varphi \rangle_{V_h', V_h} := \langle f(\mu), \varphi \rangle_{V',V} - c(\mu) \langle \nabla y_\ell(\mu), \nabla \varphi \rangle_{L^2} \\ - a(\mu) \langle \max\{0, y_\ell(\mu)\}, \varphi \rangle_{L^2} \quad \text{for all } \varphi \in V_h.\end{aligned}$$

*Proof.* Let $\mu \in \mathcal{P}$ be fixed. Then the FE solution $y_h(\mu)$ satisfies the equation

$$c(\mu) \langle \nabla y_h(\mu), \nabla \varphi \rangle_{L^2} + a(\mu) \langle \max\{0, y_h(\mu)\}, \varphi \rangle_{L^2} = \langle f(\mu), \varphi \rangle_{V',V} \tag{1.10}$$

for all $\varphi \in V_h$, and $e_y(\mu)$ is in $V_h$. We use the Poincaré inequality, the monotonicity of the max term, and (1.10) to obtain that

$$\begin{aligned}\frac{C_2}{C_1} \|e^y(\mu)\|_V^2 &\le c(\mu) \|\nabla e^y(\mu)\|_{L^2}^2 \\ &\le c(\mu) \|\nabla e^y(\mu)\|_{L^2}^2 + a(\mu) \langle \max\{0, y_h(\mu)\} - \max\{0, y_\ell(\mu)\}, y_h(\mu) - y_\ell(\mu) \rangle_{L^2} \\ &= c(\mu) \langle \nabla y_h(\mu), \nabla e^y(\mu) \rangle_{L^2} + a(\mu) \langle \max\{0, y_h(\mu)\}, e^y(\mu) \rangle_{L^2} \\ &\quad - c(\mu) \langle \nabla y_\ell(\mu), \nabla e^y(\mu) \rangle_{L^2} - a(\mu) \langle \max\{0, y_\ell(\mu)\}, e^y(\mu) \rangle_{L^2} \\ &= \langle f(\mu), e^y(\mu) \rangle_{V',V} - c(\mu) \langle \nabla y_\ell(\mu), \nabla e^y(\mu) \rangle_{L^2} - a(\mu) \langle \max\{0, y_\ell(\mu)\}, e^y(\mu) \rangle_{L^2} \\ &= \langle \mathrm{Res}_\ell^y(\mu), e^y(\mu) \rangle_{V_h', V_h} \le \|\mathrm{Res}_\ell^y(\mu)\|_{V_h'} \|e^y(\mu)\|_V,\end{aligned}$$

which gives (1.9). □

**Remark 1.2.5.** The error estimation requires us to compute the dual norm of the residual, which is equal to the primal norm of the $V$-Riesz representative of the residual. To compute the Riesz representative, we need to solve an FE discretized linear elliptic PDE. Depending on the exact application, this may end up being more costly than computing the true error, e. g., when the reduced basis of the greedy RB method ends up being large in comparison to the size of the training set $\mathcal{P}_{\text{train}}$ and the semilinear PDE can be solved with few semismooth Newton iterations.

As a direct consequence of Proposition 1.2.4, we obtain the reproduction-of-solutions and vanishing-error-bound properties, which are classical RB results.

**Corollary 1.2.6.** *Let Assumptions* 1.1.1 *and* 1.2.1 *hold. Assume that for a given* $\mu \in \mathcal{P}$, *the FE solution* $y_h(\mu) \in V_h$ *belongs to* $V_\ell$. *Then the corresponding RB solution* $y_\ell(\mu) \in V_\ell$ *satisfies*

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

1) $y_h(\mu) = y_\ell(\mu)$,
2) $\Delta_w^y(V_\ell, \mu) = 0$.

*Proof.* Part 1) follows from the proof of Proposition 1.2.4. Furthermore, we derive Part 2) directly from the definition of the residual and Part 1). □

Similarly, we can show the convergence of the RB solutions to the FE solution, as the parameters used in the reduced basis fill the parameter set.

**Theorem 1.2.7.** *Let Assumptions* 1.1.1 *and* 1.2.1 *hold, let* $(\mathcal{P}_\ell)_{\ell\in\mathbb{N}} \subset \mathcal{P}$ *be a sequence of parameter subsets, and let* $h_\ell := \sup\{\operatorname{dist}(\mu, \mathcal{P}_\ell) \mid \mu \in \mathcal{P}\}$. *Then*

$$\sup_{\mu\in\mathcal{P}} \left\|y_h(\mu) - y_\ell(\mu)\right\|_V \in \mathcal{O}(h_\ell).$$

*Proof.* For $\ell \in \mathbb{N}$ and $\mu \in \mathcal{P}$, we can choose $\mu^* \in \arg\min\{\|\mu - \tilde{\mu}\|_{\mathbb{R}^p} \mid \tilde{\mu} \in \mathcal{P}_\ell\}$. Now we use the Lipschitz continuity of the solution operator $\mathcal{S}_W$, its independence of the chosen space $W$ (Proposition 1.2.2), and the reproduction of solutions property (Corollary 1.2.6 1)) to obtain that

$$\begin{aligned}
&\left\|y_h(\mu) - y_\ell(\mu)\right\|_V \\
&\quad\le \left\|y_h(\mu) - y_h(\mu^*)\right\|_V + \left\|y_h(\mu^*) - y_\ell(\mu^*)\right\|_V + \left\|y_\ell(\mu) - y_\ell(\mu^*)\right\|_V \\
&\quad\le 2L_S \left\|\mu - \mu^*\right\|_{\mathbb{R}^p} \le 2L_S h_\ell,
\end{aligned}$$

which yields the claim. □

Hence, if the sequence $\mathcal{P}_\ell$ gets dense in $\mathcal{P}$, i. e., $\lim_{\ell\to\infty} h_\ell = 0$, then the RB solutions converge to the FE solution. As a direct consequence, we can obtain an a priori error bound.

**Corollary 1.2.8.** *Let Assumptions* 1.1.1 *and* 1.2.1 *hold, let* $\mathcal{P}_{\text{train}} \subset \mathcal{P}$ *be a discrete set with filling distance* $h_{\text{train}}$, *and let the RB greedy algorithm (Algorithm* 1.1*) terminate with tolerance* $\varepsilon_{\text{tol}}$. *Then*

$$\left\|y_h(\mu) - y_\ell(\mu)\right\|_V \le 2L_S h_{\text{train}} + \varepsilon_{\text{tol}}.$$

*Proof.* Analogous to the proof of Theorem 1.2.7. □

## 1.2.2 Numerical results for the state equation

As we have seen in Section 1.2, several classical RB results hold for the PDE in question, including convergence and error estimation results. The proofs of these results were virtually independent of the nonsmoothness property of the max operator and the solution operator to the PDE. The crucial properties were the monotonicity of the

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

max operator and the Lipschitz continuity of the solution operator to the PDE. We will confirm the results of the previous section numerically in this one and use a carefully constructed example to show that the nonsmoothness effects of the PDE are in fact the limiting factor for the RB performance in practice. Additionally, we will show a theoretical bound on the condition number of the RB system. Let us mention that the example settings introduced in this section will be reused in Section 1.3.4.

For our numerical experiments, we fix the domain $\Omega = (0,1)^2$ and consider $P_1$-type FE on a Friedrichs–Keller triangulation of the domain. The measure of fineness of the grids will be $h > 0$, which denotes the inverse number of square cells per dimension, i. e., the grid will have $2/h^2$ triangles. From here on out, we write the coefficient vector of the piecewise linear interpolant on the grid vertices of a function $z\colon \Omega \to \mathbb{R}$ in typewriter font (i. e., $\mathsf{z} \in \mathbb{R}^N$) and use the same font for the matrices in the discretized settings. Dealing with the nonlinear max term, we resort to mass lumping for this term to be able to evaluate it componentwise. Inevitably, this introduces a numerical discretization error. Its effects decrease with increasing fineness of the discretization but increase with the coefficient function $a$ that scales the nonlinearity. The corresponding stiffness matrix $\mathsf{K} \in \mathbb{R}^{N\times N}$, mass matrix $\mathsf{M} \in \mathbb{R}^{N\times N}$, lumped mass matrix $\tilde{\mathsf{M}} \in \mathbb{R}^{N\times N}$, and the right-hand side $\mathsf{f}(\mu) \in \mathbb{R}^N$ are given from the FE ansatz functions $\varphi_i$, $i = 1,\dots,N$, as

$$\mathsf{K}_{i,j} = \left(\left(\int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j \mathrm{d}\boldsymbol{x}\right)\right), \quad \mathsf{f}_i(\mu) = (\langle f(\mu), \varphi_i\rangle_{V',V}),$$
$$\mathsf{M}_{i,j} = \left(\left(\int_\Omega \varphi_i\varphi_j \mathrm{d}\boldsymbol{x}\right)\right), \quad \tilde{\mathsf{M}} = \operatorname{diag}\left(\frac{1}{3}\left|\operatorname{supp}(\varphi_i)\right| \,:\, i = 1,\dots,N\right).$$

The discretization leaves us with the following FE and RB systems:

$$c(\mu)\mathsf{K}\mathsf{y}_h(\mu) + a(\mu)\tilde{\mathsf{M}}\max\{0, \mathsf{y}_h(\mu)\} - \mathsf{f}(\mu) = 0 \quad \text{in } \mathbb{R}^N,$$
$$c(\mu)\boldsymbol{\Psi}_\ell^T\mathsf{K}\boldsymbol{\Psi}_\ell\mathsf{y}_\ell(\mu) + a(\mu)\boldsymbol{\Psi}_\ell^T\tilde{\mathsf{M}}\max\{0, \boldsymbol{\Psi}_\ell\mathsf{y}_\ell(\mu)\} - \boldsymbol{\Psi}_\ell^T\mathsf{f}(\mu) = 0 \quad \text{in } \mathbb{R}^\ell,$$

where $\boldsymbol{\Psi} \in \mathbb{R}^{N\times\ell} = [\boldsymbol{\psi}_1|\dots|\boldsymbol{\psi}_\ell]$ is the matrix whose columns are the FE coefficient vectors of the reduced basis functions. These finite-dimensional systems are solved by a standard semismooth Newton method, where the iteration matrices are

$$\mathsf{H}_h(\mathsf{y}_h(\mu)) := c(\mu)\mathsf{K} + a(\mu)\tilde{\mathsf{M}}\Theta(\mathsf{y}_h(\mu)),$$
$$\mathsf{H}_\ell(\mathsf{y}_\ell(\mu)) := c(\mu)\boldsymbol{\Psi}_\ell^T\mathsf{K}\boldsymbol{\Psi}_\ell + a(\mu)\boldsymbol{\Psi}_\ell^T\tilde{\mathsf{M}}\Theta(\boldsymbol{\Psi}_\ell\mathsf{y}_\ell(\mu))\boldsymbol{\Psi}_\ell,$$

with $\Theta := \Theta_0$, where $\Theta_x\colon \mathbb{R}^N \to \mathbb{R}^{N\times N}$ maps a vector to the diagonal matrix that takes the Heaviside function with functional value $x$ at 0 evaluated for each entry of the vector as its diagonal entries.

The RB Newton matrix possesses the favorable property that its condition number is bounded independently of the current iterate and the dimension of the RB system.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Proposition 1.2.9.** *Let* $\mathsf{Q} \in \mathbb{R}^{N\times N}$ *be symmetric and positive definite, and let* $\boldsymbol{\Psi}_\ell \in \mathbb{R}^{N\times \ell}$ *represent a reduced basis that is orthonormal with respect to the scalar product induced by* $\mathsf{Q}$*. Then the condition number of the Newton matrix* $\mathsf{H}_\ell(\mathsf{y}_\ell(\mu))$ *is bounded independently of* $\ell$*,* $y_\ell$*, and* $\mu$*.*

*Proof.* First, we fix $\ell$, $\mathsf{y}_\ell$, and $\mu$ and define $\mathsf{H} = \mathsf{H}_\ell(\mathsf{y}_\ell(\mu)) \in \mathbb{R}^{\ell\times\ell}$. Since $\mathsf{K}$ is symmetric and positive definite, the matrix $\tilde{\mathsf{M}}\Theta(\boldsymbol{\Psi}_\ell \mathsf{y}_\ell(\mu))$ is symmetric and positive semidefinite, the functions $c(\mu) > 0$, $a(\mu) \geq 0$, and $\operatorname{rank}(\boldsymbol{\Psi}_\ell) = \ell$, and the matrix $\mathsf{H}$ is symmetric and positive definite. Therefore we can compute $\operatorname{cond}_2(\mathsf{H}) = \lambda_{\max}/\lambda_{\min}$, where $\lambda_{\max} \geq \lambda_{\min} > 0$ are the largest and smallest eigenvalues, respectively, of $\mathsf{H}$. For an element $\mathsf{v} \in \mathbb{R}^\ell$ with

$$\mathsf{w} := \boldsymbol{\Psi}_\ell \mathsf{v} = \sum_{i=1}^{\ell} \mathsf{v}_i \boldsymbol{\psi}_i \in \mathbb{R}^N,$$

where $\boldsymbol{\psi}_i$ is the $i$th column of $\boldsymbol{\Psi}_\ell$, we define $\|\mathsf{w}\|_\mathsf{Q} = \langle \mathsf{Q}\mathsf{w}, \mathsf{w}\rangle_{\mathbb{R}^N}^{1/2}$. This gives

$$\|\mathsf{w}\|_\mathsf{Q}^2 = \langle \mathsf{Q}\mathsf{w}, \mathsf{w}\rangle_{\mathbb{R}^N} = \sum_{i,j=1}^{\ell} \mathsf{v}_i \mathsf{v}_j \underbrace{\langle \mathsf{Q}\boldsymbol{\psi}_i, \boldsymbol{\psi}_j\rangle_{\mathbb{R}^N}}_{=\delta_{ij}} = \sum_{i=1}^{\ell} \mathsf{v}_i^2 = \|\mathsf{v}\|_{\mathbb{R}^\ell}^2.$$

Since all finite-dimensional norms are equivalent, there are (generally, $N$-dependent) constants $\kappa_1, \kappa_2 > 0$ such that

$$\frac{1}{\kappa_1}\|\cdot\|_\mathsf{Q} \leq \|\cdot\|_\mathsf{K} \leq \kappa_1 \|\cdot\|_\mathsf{Q}, \quad \frac{1}{\kappa_2}\|\cdot\|_\mathsf{Q} \leq \|\cdot\|_{\mathbb{R}^N} \leq \kappa_2 \|\cdot\|_\mathsf{Q}.$$

Now let $\mathsf{v} \in \mathbb{R}^\ell$ be an eigenvector of $H$ for the eigenvalue $\lambda_{\max}$. We have that

$$\begin{aligned}
\lambda_{\max}\|\mathsf{v}\|_{\mathbb{R}^\ell}^2 &= \langle \mathsf{H}\mathsf{v}, \mathsf{v}\rangle_{\mathbb{R}^\ell} \\
&= c(\mu)\langle \mathsf{K}\mathsf{w}, \mathsf{w}\rangle_{\mathbb{R}^N} + a(\mu)\big\langle \tilde{\mathsf{M}}\Theta\big(\boldsymbol{\Psi}_\ell \mathsf{y}_\ell(\mu)\big)\mathsf{w}, \mathsf{w}\big\rangle_{\mathbb{R}^N} \\
&\leq c(\mu)\|\mathsf{w}\|_\mathsf{K}^2 + a(\mu)\|\tilde{\mathsf{M}}\|\|\mathsf{w}\|_{\mathbb{R}^N}^2 \leq c(\mu)\kappa_1^2\|\mathsf{w}\|_\mathsf{Q}^2 + a(\mu)\|\tilde{\mathsf{M}}\|\kappa_2^2\|\mathsf{w}\|_\mathsf{Q}^2.
\end{aligned}$$

Since $\|\mathsf{w}\|_\mathsf{Q} = \|\mathsf{v}\|_{\mathbb{R}^\ell}$, we can conclude that

$$\lambda_{\max} \leq c_1 := \kappa_1^2 \max_{\mu\in\mathcal{P}} c(\mu) + \kappa_2^2\|\tilde{\mathsf{M}}\| \max_{\mu\in\mathcal{P}} a(\mu) < \infty.$$

Now let $\mathsf{v} \in \mathbb{R}^\ell$ be an eigenvector of $\mathsf{H}$ corresponding to $\lambda_{\min}$. Then since $\tilde{\mathsf{M}}\Theta(\boldsymbol{\Psi}_\ell \mathsf{y}_\ell(\mu))$ is positive semidefinite, we can conclude that

$$\begin{aligned}
\lambda_{\min}\|\mathsf{v}\|_{\mathbb{R}^\ell}^2 = \langle \mathsf{H}\mathsf{v}, \mathsf{v}\rangle_{\mathbb{R}^\ell} &= c(\mu)\langle \mathsf{K}\mathsf{w}, \mathsf{w}\rangle_{\mathbb{R}^N} + \underbrace{a(\mu)\big\langle \tilde{\mathsf{M}}\Theta\big(\boldsymbol{\Psi}_\ell \mathsf{y}_\ell(\mu)\big)\mathsf{w}, \mathsf{w}\big\rangle_{\mathbb{R}^N}}_{\geq 0} \\
&\geq c(\mu)\langle \mathsf{K}\mathsf{w}, \mathsf{w}\rangle_{\mathbb{R}^N} \geq \frac{c(\mu)}{\kappa_1^2}\|\mathsf{w}\|_\mathsf{Q}^2.
\end{aligned}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Again, using $\|w\|_Q = \|v\|_{\mathbb{R}^\ell}$, we can conclude that

$$\lambda_{\min} \geq c_2 := \frac{\min\{c(\mu) \mid \mu \in \mathcal{P}\}}{\kappa_1^2} > 0,$$

which yields $\operatorname{cond}_2(\mathsf{H}) \leq c_1/c_2$ for $c_1$ and $c_2$ independent of $\ell$, $y_\ell$, and $\mu$. □

Our code is implemented in Python3 and uses FENICS [1] for the matrix assembly. Sparse memory management and computations are implemented with SciPy [24]. All computations below were run on an Ubuntu 18.04 notebook with 12 GB main memory and an Intel Core i7-4600U CPU.

**Example 1.2.1.** For the first numerical example, we set $\mathcal{P} = [-2, 2]^2$, $c(\mu) = 1$, $a(\mu) = 10$, and the right-hand sides $f(\mu)$ (to be understood as $L^2$-functions mapping $(x_1, x_2) \in \Omega$ to $\mathbb{R}$ embedded into $V'$) as

$$f(\mu)(x_1, x_2) = 10 \begin{cases} \mu_1 x_1 x_2 & \text{for } \boldsymbol{x} = (x_1, x_2) \text{ and } x_1 \leq \frac{1}{2}, \\ \mu_2 x_1^2 x_2^2 & \text{otherwise.} \end{cases} \tag{1.11}$$

The right-hand side and corresponding solution are shown in Figure 1.1 for two values of $\mu$. Note that the right-hand side is linear in the parameter $\mu$.

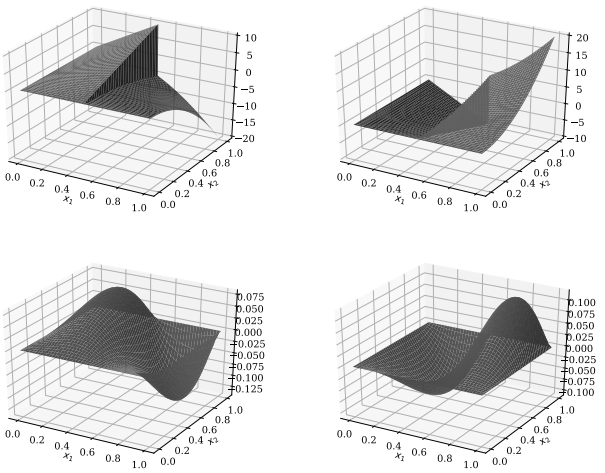


**Figure 1.1:** Example 1.2.1. Right-hand side (top) and FE solution (bottom) for parameters $\mu = (2, -2)$ (left column) and $\mu = (-2, 2)$ (right column) for $1/h = 400$.

We fix a training set of 121 and a test set of 196 equidistant points in $\mathcal{P}$, where both sets are chosen disjointly. We solve the FE system for each parameter in the test set and compare the results to those from solving the corresponding RB systems generated from offline phases that use the true $V$-error or the error estimator on the training

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

set, respectively. The semismooth Newton iterations in both FE and RB computations for a parameter are started using the solution for the closest parameter in the test set whose solution is already known as an initial value (warm starting). The semismooth Newton method terminates when the tolerance of $10^{-8}$ is reached for the dual norm of the residual and the RB error tolerance is set to $10^{-4}$. The results for different grids with step sizes $h = 1/100, 1/200, 1/400$ per dimension can be found in Table 1.1.

**Table 1.1:** Example 1.2.1. The first part of the table shows the average number of required semismooth Newton iterations for computing the FE solutions to the PDE for each parameters in the test set and the average required time. The second and third parts show the average number of semismooth Newton iterations, the average online speed-up, the required computational time for the offline phase $t_{\text{off}}$, the average true error $\|y_h - y_\ell\|_V$, and the size of the reduced basis generated in the offline phase with the true error and error estimator, respectively.

| **FE Results** | | | | | |
|---|---|---|---|---|---|
| **1/h** | **avg. iterations** | **avg. time (s)** | | | |
| 100 | 2.03 | 0.49 | | | |
| 200 | 2.03 | 3.22 | | | |
| 400 | 2.02 | 21.88 | | | |
| **offline/online RB (true $V$-error)** | | | | | |
| **1/h** | **avg. iterations** | **avg. speed-up** | $t_{\text{off}}$ (s) | avg. error | $\lvert\Psi\rvert$ |
| 100 | 2.02 | 71.00 | 70.00 | $2.38 \cdot 10^{-5}$ | 17 |
| 200 | 2.02 | 127.60 | 402.90 | $2.34 \cdot 10^{-5}$ | 17 |
| 400 | 2.02 | 250.36 | 2793.68 | $2.32 \cdot 10^{-5}$ | 17 |
| **offline/online RB (estimator)** | | | | | |
| **1/h** | **avg. iterations** | **avg. speed-up** | $t_{\text{off}}$ (s) | **avg. error** | $\lvert\Psi\rvert$ |
| 100 | 2.02 | 51.38 | 379.02 | $1.15 \cdot 10^{-5}$ | 26 |
| 200 | 2.02 | 98.60 | 2264.12 | $1.11 \cdot 10^{-5}$ | 26 |
| 400 | 2.02 | 189.65 | 16994.40 | $1.10 \cdot 10^{-5}$ | 26 |

First, we can note that all of the average $V$-errors on the test set are below the given tolerance of the offline phase that generated the reduced basis from the training set. Further, the average number of semismooth Newton iterations for solving the FE/RB systems appears to be independent of the step size $h$ and very close for the FE and RB systems. Additionally, in this example the size of the reduced basis is independent of the step size $h$ (compare with Table 1.3 for different results). Keep in mind that the very small number of about two semismooth Newton steps per system solve is due to the warm starting. Without warm starting, the average number of semismooth Newton steps per solve is around four to five iterations with the zero function as an initial guess. As the number of cells per dimension doubles in our implementation, the speed-up through RB on the test set approximately doubles as well, reaching up to a factor of 250. This means that if 130 or more online solves are required, then the RB approach including its offline cost comes out ahead of the FE approach in computational

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

time. As mentioned in Remark 1.2.5, due to the nonlinearity of the PDE, an evaluation of the error estimator means that we have to solve a linear PDE on FE level. As the number of semismooth Newton iterations per nonlinear solve is quite low in this example, it turns out that the RB approach based on the true $V$-error ends up outperforming the approach based on the error estimator in the offline phase and, as it works on a smaller reduced basis, in the online phase as well. Nonetheless, the error estimator is useful to estimate the error in the online phase as shown in Table 1.2. The effectivity (the quotient of the estimated and true errors) appears to be mesh independent. The speed-up is approximately four. This independence of the step size can be expected, since the main difference between the true and estimated errors is whether a linear or nonlinear PDE needs to be solved on the FE level. Again, a larger speed-up cannot be achieved, since the error estimator cannot be implemented in a parameter separable fashion. See also Remark 1.2.5.

**Table 1.2:** Example 1.2.1. Performance of true error and error estimator with reduced basis from strong greedy on test set.

| 1/h | avg. true $V$-error | avg. effectivity | time true $V$-error (s) | speed-up estimator |
|---|---|---|---|---|
| 100 | $2.38 \cdot 10^{-5}$ | 1.97 | 96.69 | 4.32 |
| 200 | $2.34 \cdot 10^{-5}$ | 1.99 | 628.16 | 4.10 |
| 400 | $2.32 \cdot 10^{-5}$ | 1.99 | 4242.60 | 4.10 |

**Example 1.2.2.** In the second example, we set $P = [0,3]$, $c(\mu) = 1$, and $a(\mu) = 8\pi^2\mu$. In contrast to Example 1, the right-hand side is nonlinear in the parameter, and we ensure that nonsmoothness effects occur by constructing a corresponding known solution that passes through zero on nonnegligible parts of the domain as the parameter $\mu$ varies. We construct an appropriate right-hand side by applying the partial differential operator to the solution. To that end, we denote the four quarters of our domain as

$$\Omega_0 := \left[0, \frac{1}{2}\right] \times \left[0, \frac{1}{2}\right], \quad \Omega_1 := \left[\frac{1}{2}, 1\right] \times \left[0, \frac{1}{2}\right],$$
$$\Omega_2 := \left[\frac{1}{2}, 1\right] \times \left[\frac{1}{2}, 1\right], \quad \Omega_3 := \left[0, \frac{1}{2}\right] \times \left[\frac{1}{2}, 1\right],$$

and define

$$\begin{aligned} g &: \mathbb{R}^2 \to \mathbb{R}, \quad \boldsymbol{x} = (x_1, x_2) \mapsto \sin^2(2\pi x_1)\sin^2(2\pi x_2), \\ b_1 &: \mathcal{P} \to \mathbb{R}, \quad \mu \mapsto 1 - 2(\min\{\max\{\mu, 2\}, 3\} - 2), \\ b_2 &: \mathcal{P} \to \mathbb{R}, \quad \mu \mapsto -1 + 2(\min\{\max\{\mu, 1\}, 2\} - 1), \\ b_3 &: \mathcal{P} \to \mathbb{R}, \quad \mu \mapsto 1 - 2\min\{\max\{\mu, 0\}, 1\}. \end{aligned}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

We construct the sought out analytical solution of the PDE as

$$y(\mu)\colon \Omega \to \mathbb{R}, \quad \boldsymbol{x} = (x_1, x_2) \mapsto \begin{cases} 0, & \boldsymbol{x} \in \Omega_0, \\ b_1(\mu)g(\boldsymbol{x}), & \boldsymbol{x} \in \Omega_1, \\ b_2(\mu)g(\boldsymbol{x}), & \boldsymbol{x} \in \Omega_2, \\ b_3(\mu)g(\boldsymbol{x}), & \boldsymbol{x} \in \Omega_3. \end{cases} \tag{1.12}$$

The solution is designed as nonconstant in three quarters of the domain. We divide the parameter space into three equal parts as $\mathcal{P} = [0,3] = [0,1] \cup [1,2] \cup [2,3]$. As the parameter passes through either of the thirds of the parameter space, it flips the sign of a "bump" function with support in one of the quarters, which is prescribed by $g$. The change in the sign is given by the piecewise linear functions $b_i$. Thus the amplitude of the bump changes linearly as well. Accordingly, the solution passes through 0 on an entire quarter of the domain whenever the parameter hits the middle of either of the thirds of the parameter interval. The solution is twice continuously differentiable in space, and therefore the corresponding right-hand side is given by $f(\mu) := -\Delta y(\mu) + 8\pi^2 \mu \max\{0, y(\mu)\}$, where the Laplacian can be taken in the strong sense. The solution for some parameters and the functions $b_i$ can be seen in Figure 1.2. For this example, we choose a training set of 60 and a disjoint test set of 100 equidistant points in $\mathcal{P} = [0,3]$. The semismooth Newton iterations in both FE and RB computations for a parameter are warm started as previously. The semismooth Newton method terminates when the tolerance of $10^{-8}$ is reached for the dual norm of the residual and the RB error tolerance is set to $10^{-4}$. The results for different grids with step sizes $h = 1/100, 1/200, 1/400$ per dimension can be found in Table 1.3. As in Example 1.2.1, the average RB error on the test set is below the tolerance imposed on the training set. Additionally, the average number of semismooth Newton iterations appears to be mesh independent in this example as well. Again, the low number of average iterations is due to warm starting. As the number of cells in the grid increases, the number of functions in the reduced basis decreases in this example. This effect is unsurprising since the discretization captures the effects of the continuous PDE more accurately with increased fineness of the grids. As we know the solution, we expect three inde-

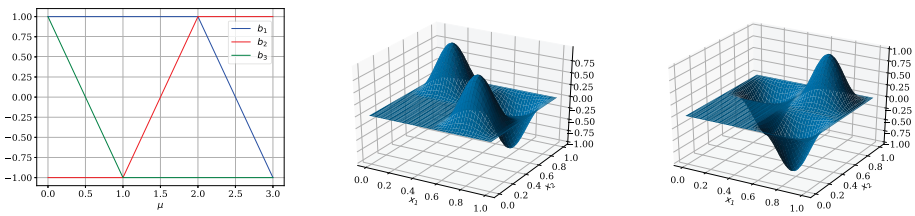



**Figure 1.2:** Example 1.2.2. Functions $b_i$ (left) and FE solution for parameters $\mu = 0$ (middle) and $\mu = 3$ (right) for $1/h = 400$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Table 1.3:** Example 1.2.2. The first part of the table shows the average number of required semismooth Newton iterations for computing the FE solutions to the PDE for each parameter in the test set and the average required time. The second and third parts show the average numbers of semismooth Newton iterations, the average online speed-up, the required computational time for the offline phase $t_{\text{off}}$, the average error $\|y_h - y_\ell\|_V$, and the size of the reduced basis generated in the offline phase with the true and estimated errors, respectively.

| **FE Results** | | | | | |
|---|---|---|---|---|---|
| **1/h** | **avg. iterations** | **avg. time (s)** | | | |
| 100 | 2.45 | 0.52 | | | |
| 200 | 2.41 | 3.51 | | | |
| 400 | 2.31 | 24.02 | | | |
| **offline/online RB (true $V$-error)** | | | | | |
| **1/h** | **avg. iterations** | **avg. speed-up** | $t_{\text{off}}$ (s) | **avg. error** | $\lvert\Psi\rvert$ |
| 100 | 2.40 | 57.54 | 43.57 | $4.81 \cdot 10^{-5}$ | 24 |
| 200 | 2.32 | 131.92 | 266.68 | $3.98 \cdot 10^{-5}$ | 19 |
| 400 | 2.26 | 378.26 | 1763.79 | $2.80 \cdot 10^{-5}$ | 16 |
| **offline/online RB (estimator)** | | | | | |
| **1/h** | **avg. iterations** | **avg. speed-up** | $t_{\text{off}}$ (s) | **avg. error** | $\lvert\Psi\rvert$ |
| 100 | 2.39 | 47.26 | 177.29 | $3.64 \cdot 10^{-5}$ | 29 |
| 200 | 2.37 | 115.46 | 1050.11 | $2.08 \cdot 10^{-5}$ | 24 |
| 400 | 2.28 | 220.90 | 6427.29 | $1.46 \cdot 10^{-5}$ | 19 |

**Table 1.4:** Example 1.2.2. Performance of true error and error estimator with reduced basis from strong greedy on test set.

| **1/h** | **avg. true $V$-error** | **avg. effectivity** | **time true $V$-error (s)** | **speed-up estimator** |
|---|---|---|---|---|
| 100 | $4.81 \cdot 19^{-5}$ | 2.08 | 66.93 | 2.66 |
| 200 | $3.98 \cdot 10^{-5}$ | 2.08 | 384.70 | 3.43 |
| 400 | $2.80 \cdot 10^{-5}$ | 2.13 | 2489.67 | 3.78 |

pendent parts, i. e., an RB with at least three functions. When the discretization is taken to have 1024 square cells per dimension, the RB computed using the true error actually reaches this "lower bound". We do not include the full computational results for this discretization here because the computation was run on a different system, so there is no comparability. The speed-up can be observed to roughly double as the number of cells per spatial dimension doubles as well, similarly to Example 1.2.1. Also, as in Example 1.2.1, we can observe that the reduced basis approach with the true error outperforms the approach using the error estimator. The favorable effectivity of the error estimator on the test set carries over from Example 1.2.1 (see Table 1.4), and the speed-up factor seems to be slightly below four as well, though the factor is not quite as stable for coarser grids in comparison to Example 1.2.1. This might be due to the nonconstant size of the reduced basis. Recall that the constructed solution $y(\mu)$ to the PDE in this example has three bumps, which are flipped with the change of sign oc-

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

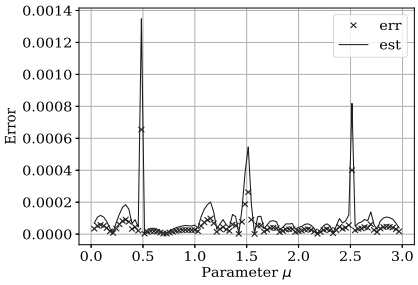



**Figure 1.3:** Example 1.2.2. True error and error estimate over the test set $\mathcal{P}_{\text{test}}$ with reduced basis from strong greedy on test set for $1/h = 400$.

curring at 0.5, 1.5, and 2.5, that is, the midpoints of the intervals $[0,1]$, $[1,2]$, and $[2,3]$ that the parameter space splits up into. These are the points where the nonlinearity and nondifferentiability of the max operator really influences the behavior of the solutions. Looking at Figure 1.3, we can observe that the true error and the error estimate over the test set are highest at these points of nondifferentiability, which suggests that reducing the information of the nondifferentiability in the PDE is the dominating task in the model order reduction.

## 1.3 MOR for the optimal control problem

This section addresses the RB-based model order reduction for the PDE constrained optimal control problem (**P**). Specifically, we combine the pseudo-semismooth Newton approach employed in [8, 9] with reduced basis techniques applied to the state $\bar{y}$ and the adjoint state $\bar{p}$ for solving the first-order system

$$\Phi(\bar{y},\bar{p},\bar{\mu}) := \begin{pmatrix} -\Delta\bar{y} + \max\{0,\bar{y}\} - \mathcal{B}\bar{\mu} \\ -\Delta\bar{p} + \mathbb{1}_{\{\bar{y}>0\}}\bar{p} - j'(\bar{y}) \\ \mathcal{B}^*\bar{p} + \sigma\mathsf{A}\mu \end{pmatrix} = 0, \tag{1.13}$$

which is to be understood as a problem in $V' \times V' \times \mathbb{R}^p$. The PSN iterations ignore the nondifferentiable indicator function in the linearization of the residual. Therefore the system matrix for this system at an iterate $(y,p,\mu)$ reads as

$$\begin{pmatrix} -\Delta + \mathbb{1}_{\{y>0\}} & 0 & -\mathcal{B} \\ -j''(y) & -\Delta + \mathbb{1}_{\{y>0\}} & 0 \\ 0 & \mathcal{B}^* & \sigma\mathsf{A} \end{pmatrix}.$$

If distributed controls in $L^2(\Omega)$ are considered in the optimization problem (**P**) with $\mathcal{B}$ being the canonical embedding $L^2(\Omega) \hookrightarrow H^{-1}(\Omega)$ and $\|\cdot\|_{\mathsf{A}}$ is replaced by $\|\cdot\|_{L^2}$, then (1.13) is a strong necessary first-order system for minimizers of (**P**), as shown in [9], in the sense that it is equivalent to the variational inequality characterizing

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

purely primal first-order stationarity. Given our setting of Assumption 1.1.1 and finite-dimensional controls in (**P**), solutions to (1.13) are generally stationary in a slightly weaker sense (see also [8]).

We derive an error estimator and indicator for the primal and the adjoint state in (1.13). Further, we present a classical offline/online RB approach and a novel adaptive RB approach that combines the PSN iterations and online updates of the reduced basis. This section should be understood as a numerical study comparing both approaches, and analytical results will be limited to the error estimator and indicator.

Note that contrary to the assumptions made in the previous section, we do not assume any constraints on the parameter set $\mathcal{P} = \mathbb{R}^p$ in this section, the reason being that while we can identify a parameter region of interest in the forward problem beforehand, it is quite unclear where the optimal parameter $\mu$ in the optimization problem will be located. However, since $\mathcal{P}$ is nonempty and closed, the setting of the first-order system state equation satisfies Assumption 1.2.1 and is therefore covered by the analysis in the previous section.

**Remark 1.3.1.** In [9], system (1.13) is reduced to the state and the adjoint state by eliminating the parameter using the last line in the system. However, the reduced system involves the operator $\mathcal{B}\mathcal{B}^*$, whose discretized matrix form is dense, introducing memory issues for our algorithms, which use sparse matrix structure when possible, in practice. Hence we will keep the original three-line system (1.13) in the state, adjoint, and parameter as is.

### 1.3.1 Offline/online RB approach

The classical offline/online RB approach is similar to that in Section 1.2. As mentioned above, here we are dealing with $\mathcal{P} = \mathbb{R}^p$, but the offline phase of the standard offline/online RB approach requires a discrete, compact training set to generate the reduced basis. In practice, we will therefore use a discrete subset of a heuristically fixed box centered at the origin that is "sufficiently large" in some sense, i. e., by including all PSN iterates and the optimal parameter, as the training set $\mathcal{P}_{\text{train}} \subset \mathbb{R}^p$.

**Remark 1.3.2.** When the state-dependent part of the cost functional $j$ is bounded from below by $\xi$, it is possible to compute a compact set of parameters that includes all possible optimal parameters. Starting with an arbitrary $\mu_0 \in \mathbb{R}^p$, we have that

$$\mathcal{J}(y(\mu), \mu) \geq \xi + \frac{\sigma}{2}\|\mu\|_{\mathsf{A}}^2 > \mathcal{J}(y(\mu_0), \mu_0)$$

for all $\mu \in \mathbb{R}^p$ with $\|\mu\|_{\mathsf{A}}^2 > 2(\mathcal{J}(y(\mu_0), \mu_0) - \xi)/\sigma$. Thus we can choose the set $\{\mu \in \mathbb{R}^p : \|\mu\|_{\mathsf{A}}^2 \leq 2(\mathcal{J}(y(\mu_0), \mu_0) - \xi)/\sigma\}$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

We use a common reduced basis $\Psi_\ell$ and reduced space $V_\ell$ for state and adjoint. Accordingly, we end up with Algorithm 1.2 for the offline phase of the reduced basis method in the optimal control problem.

**Algorithm 1.2:** Greedy RB method for first-order system.

**Require:** Discrete training set of parameters $\mathcal{P}_{\text{train}} \subset \mathcal{P}$, error tolerance $\varepsilon_{\text{tol}} > 0$
**Return :** RB parameters $\mathcal{P}_\ell$, reduced basis $\Psi_\ell$, RB space $V_\ell$
Set $\ell = 0$, $\mathcal{P}_0 = \emptyset$, $\Psi_0 = \emptyset$, $V_0 = \{0\}$;
**while** $\varepsilon_\ell := \max\{\Delta(V_\ell, \mu) \,|\, \mu \in \mathcal{P}_{\text{train}}\} > \varepsilon_{\text{tol}}$ **do**
  Compute $\mu_{\ell+1} \in \arg\max\{\Delta(V_\ell, \mu) \,|\, \mu \in \mathcal{P}_{\text{train}}\}$;
  Set $\mathcal{P}_{\ell+1} = \mathcal{P}_\ell \cup \{\mu_{\ell+1}\}$, $\psi^y_{\ell+1} = y_h(\mu_{\ell+1})$ and $\psi^p_{\ell+1} = p_h(\mu_{\ell+1})$;
  Orthonormalize $(\psi^y_{\ell+1}, \psi^p_{\ell+1})$ against $\Psi_\ell$;
  Set $\Psi_{\ell+1} = \Psi_\ell \cup (\{\psi^y_{\ell+1}, \psi^p_{\ell+1}\} \setminus \{0\})$;
  Define $V_{\ell+1} = V_\ell \oplus \operatorname{span}(\psi^y_{\ell+1}, \psi^p_{\ell+1})$ and $\ell = \ell + 1$;

Again, the symbol $\Delta(V_\ell, \mu)$ denotes either the true error or an error indicator with respect to a given RB space $V_\ell$. Since we apply the basis reduction on the state and the adjoint, the true error combines both errors, i. e.,

$$e_\ell(\mu) = \begin{pmatrix} e_y(\mu) \\ e_p(\mu) \end{pmatrix}_\ell := \begin{pmatrix} y_h(\mu) - y_\ell(\mu) \\ p_h(\mu) - p_\ell(\mu) \end{pmatrix},$$
$$\|e_\ell(\mu)\| := \|y_h(\mu) - y_\ell(\mu)\|_V + \|p_h(\mu) - p_\ell(\mu)\|_V.$$

A residual-based error indicator will be addressed in the section after next. Once the offline phase is completed, we solve the FE/RB discretized version of the first-order system (1.13) with the PSN approach.

## 1.3.2 Adaptive RB approach

As any of the classical offline/online greedy RB approaches, the method described in the previous subsection suffers from a number of drawbacks. The curse of dimensionality leads to an exponential growth in the offline phase computation time, which is detrimental for high-dimensional parameter spaces. This means that even if the online speed-up is significant, the computational effort of the offline phase can make the overall computational time of the RB approach exceed that of the standard FE problem. Additionally, choosing a compact training set that includes the optimal parameter for the offline phase is nontrivial. These drawbacks are avoided in the adaptive RB

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

approach that we propose in this section. The method starts out with an initial starting guess for the solution and alternates PSN iterations and RB updates to improve the quality of the approximations of FE elements by RB elements and of the FE solutions with the PSN iterates using local information, where local refers to the information at each of the iterates the PSN algorithm reaches. Since all of the model order reduction work is taken care of along the way, it is clear that this approach is geared toward accelerating a single solve of the first-order system without the added expense of an offline phase – it is quite complementary to the offline/online approach in that sense. When only few, but more that one, solves of the first-order system are required, e. g., for varying the Tikhonov parameter $\sigma$ in the cost functional, reusing the basis generated in the first solve is of course an option that will likely outperform offline/online approaches due to their large offline cost. The adaptive approach is somewhat similar to that in [19], where trust region methods are combined with an adaptive RB approach. Our complete algorithm is described in Algorithm 1.3.

**Algorithm 1.3:** Adaptive RB method for first-order system.

**Require:** Initial parameter $\mu_0 \in \mathbb{R}^p$, error tolerances $\varepsilon_N, \varepsilon_{\mathrm{RB}} > 0$ for RB and PSN, weight $1 > \eta > 0$, $n_{\mathrm{fix}} \geq 0$

**Return :** RB approximation $(y_n, p_n, \mu_n)$ of FE solution $(y_h, p_h, \mu_h)$

Set $\ell = 0$ (number of basis updates);

Set $n = 0$ (number of PSN iterations);

Set $\mathcal{P}_0 = \emptyset$, $\Psi_0 = \emptyset$, $V_0 = \emptyset$;

Set $y_0 = y_h(\mu_0)$, $p_0 = p_h(\mu_0)$;

**while** $\Delta(V_\ell, \mu_n) > \varepsilon_{\mathrm{RB}}$ *or* $\|\Phi(y_n, p_n, \mu_n)\| > \varepsilon_N$ **do**

  **if** $\ell > 0$ *and* $\frac{(\Delta(V_\ell,\mu_n)-\varepsilon_{\mathrm{RB}})_+}{\varepsilon_{\mathrm{RB}}} < \eta^n \frac{(\|\Phi(y_n,p_n,\mu_n)\|-\varepsilon_N)_+}{\varepsilon_N}$ **then**

    Get $(y_{n+1}, p_{n+1}, \mu_{n+1})$ from PSN step w. r. t. $\Psi_\ell$;

    Set $n = n + 1$;

  **else**

    Compute $\psi^y_{\ell+1} = y(\mu_n)$ and $\psi^p_{\ell+1} = p(\mu_n)$;

    Orthonormalize $(\psi^y_{\ell+1}, \psi^p_{\ell+1})$ against $\Psi_\ell$;

    Set $\mathcal{P}_{\ell+1} = \mathcal{P}_\ell \cup \{\mu_n\}$, $\Psi_{\ell+1} = \Psi_\ell \cup (\{\psi^y_{\ell+1}, \psi^p_{\ell+1}\} \setminus \{0\})$, $V_{\ell+1} = V_\ell \oplus \operatorname{span}(\psi^y_{\ell+1}, \psi^p_{\ell+1})$;

    Set $\ell = \ell + 1$;

    Get $(y_{n+n_{\mathrm{fix}}}, p_{n+n_{\mathrm{fix}}}, \mu_{n+n_{\mathrm{fix}}})$ from $n_{\mathrm{fix}}$ PSN steps w. r. t. $\Psi_\ell$;

    Set $n = n + n_{\mathrm{fix}}$;

The algorithm exploits a balancing of the fact that the PSN residual needs to be small only if the RB approximation is sufficiently accurate and vice versa. Starting from the

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

initial guess, the algorithm repeatedly checks whether the RB or the PSN tolerance is violated. If so, then the relative violation of the tolerances is compared, and depending on their magnitude, either further PSN iterations are applied, or the basis is updated. We include a biasing factor $\eta^n$ with $0 < \eta < 1$ in the comparison of the violations, which ensures that an improvement of the RB approximation quality is favored over additional PSN steps as the number of PSN iterations increases. In the early stages of the PSN iterations, where the residual is large and we are comparatively far from a solution and PSN updates are relatively coarse, the algorithm therefore benefits from small system sizes due to the small reduced bases. With increasing proximity to a solution, the RB approximation quality is increased to avoid pointless PSN steps on systems that do not capture the dynamics of the solution sufficiently well. By updating the reduced basis locally, i. e., along the PSN iterates, which are expected to exhibit behavior similar to that of the solution, we increase the quality of the RB approximation of the solution compared to a "one-size-fits-all" type offline/online approach.

Since the evaluation of the error indicator and basis updates are computationally expensive themselves and increase the computational cost in each of the system solves of the PSN steps, we perform a fixed number $n_{\text{fix}}$ of PSN steps whenever the reduced basis is updated to ensure that some progress toward the solution is made before the basis is expanded. This is of course an additional parameter that requires dialing in, but the numerical results in Section 1.3.4 suggest that choosing this parameter greater than one can indeed be beneficial. Note that whereas enlarging the reduced basis increases the system size, both the basis updates and the evaluation of the true error usually gain efficiency as the iterations progress because current RB iterates (embedded into the FE space) can be used as initial guesses for the solves that have to be performed on the FE level. In practice, we will see that use of the true $V$-error can outperform an indicator derived below, which is in part due to this behavior.

### 1.3.3 Error estimator and indicator

Error estimators for the RB approximation of the first-order system need to address the errors in the state and adjoint, i. e., $\|y_h(\mu) - y_\ell(\mu)\|_V$ and $\|p_h(\mu) - p_\ell(\mu)\|_V$. Of course, the more general results for the state in Proposition 1.2.4 that we derived in Section 1.2 can be directly transferred to this setting. For the adjoint state, the situation is more complicated. Due to the discontinuous indicator function in the adjoint equation, straightforward estimation is quite limited. The primary result based on the Lipschitz continuity of the cost functional is summarized in the following lemma.

**Lemma 1.3.3.** *Let Assumption* 1.1.1 *hold, and let $j'$ be Lipschitz continuous with constant $\gamma > 0$. Then*

$$\left\|p_h(\mu) - p_\ell(\mu)\right\|_V \leq C_1\left(\gamma\Delta_w^\gamma(V_\ell, \mu) + \left\|\mathrm{Res}_\ell^p(\mu)\right\|_{V_h'} + \left\|p_\ell(\mu)\right\|_{L^2}\right)$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*for all $\mu \in \mathcal{P}$, where $C_P$ is the Poincaré constant of the domain, $C_1 = 1 + C_P^2$, and the adjoint residual $\mathrm{Res}_\ell^p \colon \mathcal{P} \to V_h'$ is defined as*

$$\left\langle \mathrm{Res}_\ell^p(\mu), \varphi \right\rangle_{V_h', V_h} := \left\langle j'(y_\ell(\mu)), \varphi \right\rangle_{V',V} - \left\langle \nabla p_\ell(\mu), \nabla \varphi \right\rangle_{L^2} - \left\langle \mathbb{1}_{\{y_\ell(\mu)>0\}} p_\ell(\mu), \varphi \right\rangle_{L^2}$$

*for $\varphi \in V_h$.*

*Proof.* With $e_p(\mu) = p_h(\mu) - p_\ell(\mu)$, the Poincaré inequality implies that

$$\begin{aligned}
\frac{\|e_p(\mu)\|_V^2}{C_1} &\le \left\|\nabla(e_p(\mu))\right\|_{L^2}^2 \\
&\le \left\|\nabla e_p(\mu)\right\|_{L^2}^2 + \left\langle \mathbb{1}_{\{y_h(\mu)>0\}} e_p(\mu), e_p(\mu) \right\rangle_{L^2} \\
&= \left\langle j'(y_h(\mu)), e_p(\mu) \right\rangle_{V',V} - \left\langle \nabla p_\ell(\mu), \nabla e_p(\mu) \right\rangle_{L^2} \\
&\quad - \left\langle \mathbb{1}_{\{y_h(\mu)>0\}} p_\ell(\mu), e_p(\mu) \right\rangle_{L^2},
\end{aligned}$$

where we have used the adjoint equation in the last equality. Now we utilize the assumption on $j'$ and Proposition 1.2.4 to conclude that

$$\begin{aligned}
\frac{\|e_p(\mu)\|_V^2}{C_1} &\le \left\langle j'(y_h(\mu)) - j'(y_\ell(\mu)), e_p(\mu) \right\rangle_{V',V} + \left\langle j'(y_\ell(\mu)), e_p(\mu) \right\rangle_{V',V} \\
&\quad - \left\langle \nabla p_\ell(\mu), \nabla e_p(\mu) \right\rangle_{L^2} - \left\langle \mathbb{1}_{\{y(\mu)>0\}} p_\ell(\mu), e_p(\mu) \right\rangle_{L^2} \\
&\le \gamma \|e_y(\mu)\|_V \|e_p(\mu)\|_V + \left\langle j'(y_\ell(\mu)), e_p(\mu) \right\rangle_{V',V} \\
&\quad - \left\langle \nabla p_\ell(\mu), \nabla e_p(\mu) \right\rangle_{L^2} - \left\langle \mathbb{1}_{\{y_\ell(\mu)>0\}} p_\ell(\mu), e_p(\mu) \right\rangle_{L^2} \\
&\quad + \left\langle (\mathbb{1}_{\{y_\ell(\mu)>0\}} - \mathbb{1}_{\{y_h(\mu)>0\}}) p_\ell(\mu), e_p(\mu) \right\rangle_{L^2} \\
&\le (\gamma C \|\mathrm{Res}_\ell^y(\mu)\|_{V_h'} + \|\mathrm{Res}_\ell^p(\mu)\|_{V_h'} \\
&\quad + \|(\mathbb{1}_{\{y_\ell(\mu)>0\}} - \mathbb{1}_{\{y_h(\mu)>0\}}) p_\ell(\mu)\|_{L^2}) \|e_p(\mu)\|_V.
\end{aligned}$$

This especially implies that

$$\begin{aligned}
\|p_h(\mu) - p_\ell(\mu)\|_V \le C_1 (\gamma \Delta_w^y(V_\ell, \mu) &+ \|\mathrm{Res}_\ell^p(\mu)\|_{V_h'} \\
&+ \|(\mathbb{1}_{\{y_\ell(\mu)>0\}} - \mathbb{1}_{\{y_h(\mu)>0\}}) p_\ell(\mu)\|_{L^2}),
\end{aligned} \tag{1.14}$$

and the boundedness of the absolute value of the difference of the indicator functions yields the claim. □

The obvious problem with this bound is the dependence of the RB adjoint state on the $L^2$-error, which can cause overestimation of the true adjoint error even if we have the convergence of the RB to the FE solution. The reason that we get stuck with this undesirable term is that we estimate the difference of the indicator functions in (1.14)

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

very coarsely by its $L^\infty$-upper bound. The difference of the indicator functions is intrinsically difficult to handle analytically because of their discontinuous behavior and the fact that the difference multiplied by an FE function generally is no longer in the FE space. Note, however, that using the previous result numerically, we would depend on quadrature rules to actually compute the $L^2$-norm term; in our case, this would be realized by the mass lumping approach. It turns out that we can improve on the results presented in the previous lemma when we work in this mass lumped setting. To that end, we introduce some additional notation. Since the notation for functions and their coefficient vectors are inevitably going to mix in the following, recall that we introduced the typewriter font for coefficient vectors in Subsection 1.2.2. For the product of an indicator function and a $V_h$-function $z$, we let $\varepsilon_{\tilde{\mathtt{M}}}(z) := |\|z\|_{L^2} - (\mathtt{z}^T\tilde{\mathtt{M}}\mathtt{z})^{\frac{1}{2}}|$ denote the error introduced by replacing the $L^2$-norm with the mass lumping quadrature. Note that, of course, this is a grid-dependent quantity. Additionally, we introduce the shorthand notation

$$\mathtt{g}^{\pm}_{\varepsilon_{\tilde{\mathtt{M}}}}(\mu) := \mathbf{\Psi}_\ell \mathtt{y}_\ell(\mu) \pm (\Delta^y_w(V_\ell,\mu) + \varepsilon_{\tilde{\mathtt{M}}}(y_h(\mu) - y_\ell(\mu)))\tilde{\mathtt{M}}^{-\frac{1}{2}}\mathbf{1} \in \mathbb{R}^N, \tag{1.15}$$

$$\mathtt{g}^{\pm}(\mu) := \mathbf{\Psi}_\ell \mathtt{y}_\ell(\mu) \pm (\Delta^y_w(V_\ell,\mu))\tilde{\mathtt{M}}^{-\frac{1}{2}}\mathbf{1} \in \mathbb{R}^N, \tag{1.16}$$

where $\mathbf{1} \in \mathbb{R}^N$ denotes the vector of all ones. For improved readability, we will notationally suppress the dependencies of functions and vectors on the parameter $\mu$ for the remainder of this subsection. With these quantities, we can obtain the following estimate that is based on the discrete implementation. Note that the only remaining dependence on any FE function in the right-hand side of this statement is in the mass lumping error. We will disregard this dependence further down the line to obtain a reasonable (mesh-dependent) error indicator.

**Lemma 1.3.4.** *Let Assumption* 1.1.1 *hold, let $j'$ be Lipschitz continuous with constant $\gamma > 0$, and let $C_1 = 1 + C_P^2$, where $C_P$ is the Poincaré constant. Then*

$$\left\|p_h(\mu) - p_\ell(\mu)\right\|_V \leq \Delta^p_{w,\varepsilon_{\tilde{\mathtt{M}}}}(V_\ell,\mu) \quad \textit{for every } \mu \in \mathcal{P}$$

*with the term*

$$\begin{aligned}\Delta^p_{w,\varepsilon_{\tilde{\mathtt{M}}}}(V_\ell,\mu) := C_1\big(&\gamma\Delta^y_w(V_\ell,\mu) + \left\|\mathrm{Res}^p_\ell(\mu)\right\|_{V_h'} + \varepsilon_{\tilde{\mathtt{M}}}\big((\mathbb{1}_{\{y_\ell>0\}} - \mathbb{1}_{\{y_h>0\}})p_\ell\big)\\ &+ \big[\big(\Theta(\mathtt{g}^+_{\varepsilon_{\tilde{\mathtt{M}}}})\mathbf{\Psi}_\ell \mathtt{p}_\ell\big)^T\tilde{\mathtt{M}}\Theta_1(-\mathbf{\Psi}_\ell \mathtt{y}_\ell)\big(\Theta(\mathtt{g}^+_{\varepsilon_{\tilde{\mathtt{M}}}})\mathbf{\Psi}_\ell \mathtt{p}_\ell\big)\\ &+ \big(\big(\Theta(\mathbf{\Psi}_\ell \mathtt{y}_\ell) - \Theta(\mathtt{g}^-_{\varepsilon_{\tilde{\mathtt{M}}}})\big)\mathbf{\Psi}_\ell \mathtt{p}_\ell\big)^T\tilde{\mathtt{M}}\Theta(\mathbf{\Psi}_\ell \mathtt{y}_\ell)\big(\big(\Theta(\mathbf{\Psi}_\ell \mathtt{y}_\ell) - \Theta(\mathtt{g}^-_{\varepsilon_{\tilde{\mathtt{M}}}})\big)\mathbf{\Psi}_\ell \mathtt{p}_\ell\big)\big]^{\frac{1}{2}}\big)\end{aligned}$$

*and $\Theta = \Theta_0$ with $\Theta_x$ as defined in Section* 1.2.2.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*Proof.* From (1.14), replacing the exact $L^2$-norm by the mass lumped computation, we immediately obtain that

$$\begin{aligned}\|p_h(\mu) - p_\ell(\mu)\|_V \le C_1\big(&\gamma\Delta_w^y(V_\ell,\mu) + \|\mathrm{Res}_\ell^p(\mu)\|_{V_h'} \\ &+ \big[\big((\Theta(\mathbf{\Psi}_\ell \mathsf{y}_\ell) - \Theta(\mathsf{y}_h))\mathbf{\Psi}_\ell \mathsf{p}_\ell\big)^T \tilde{\mathsf{M}}\big((\Theta(\mathbf{\Psi}_\ell \mathsf{y}_\ell) - \Theta(\mathsf{y}_h))\mathbf{\Psi}_\ell \mathsf{p}_\ell\big)\big]^{\frac{1}{2}} \\ &+ \varepsilon_{\tilde{\mathsf{M}}}\big((\mathbb{1}_{\{y_\ell>0\}} - \mathbb{1}_{\{y_h>0\}})p_\ell\big)\big).\end{aligned}$$

We split up the mass lumped norm term in the middle into the parts corresponding to the index sets of the nonpositive and positive nodal values of the RB state embedded into the FE space, which gives

$$\begin{aligned}&\big((\Theta(\mathbf{\Psi}_\ell \mathsf{y}_\ell) - \Theta(\mathsf{y}_h))\mathbf{\Psi}_\ell \mathsf{p}_\ell\big)^T \tilde{\mathsf{M}}\big((\Theta(\mathbf{\Psi}_\ell \mathsf{y}_\ell) - \Theta(\mathsf{y}_h))\mathbf{\Psi}_\ell \mathsf{p}_\ell\big) \\ &\quad = \big(\Theta(\mathsf{y}_h)\mathbf{\Psi}_\ell \mathsf{p}_\ell\big)^T \tilde{\mathsf{M}}\Theta_1(-\mathbf{\Psi}_\ell \mathsf{y}_\ell)\big(\Theta(\mathsf{y}_h)\mathbf{\Psi}_\ell \mathsf{p}_\ell\big) \\ &\qquad + \big((\Theta(\mathbf{\Psi}_\ell \mathsf{y}_\ell) - \Theta(\mathsf{y}_h))\mathbf{\Psi}_\ell \mathsf{p}_\ell\big)^T \tilde{\mathsf{M}}\big(\Theta(\mathbf{\Psi}_\ell \mathsf{y}_\ell)\big)\big((\Theta(\mathbf{\Psi}_\ell \mathsf{y}_\ell) - \Theta(\mathsf{y}_h))\mathbf{\Psi}_\ell \mathsf{p}_\ell\big).\end{aligned}$$

To get rid of the dependency on the FE solution $\mathsf{y}_h$, we derive componentwise upper and lower bounds that depend on RB quantities only. Using the quadrature error and the state error estimate in Proposition 1.2.4, we obtain that

$$\begin{aligned}\big[(\mathsf{y}_h - \mathbf{\Psi}_\ell \mathsf{y}_\ell)^T \tilde{\mathsf{M}}(\mathsf{y}_h - \mathbf{\Psi}_\ell \mathsf{y}_\ell)\big]^{\frac{1}{2}} &\le \|(y_h - y_\ell)\|_V + \varepsilon_{\tilde{\mathsf{M}}}(y_h - y_\ell) \\ &\le \Delta_w^y(V_\ell,\mu) + \varepsilon_{\tilde{\mathsf{M}}}(y_h - y_\ell) \\ &= C_1\|\mathrm{Res}_\ell^y(\mu)\|_{V_h'} + \varepsilon_{\tilde{\mathsf{M}}}(y_h - y_\ell),\end{aligned}$$

and since $\tilde{\mathsf{M}}$ is a positive definite diagonal matrix, we gather that

$$\big|(\mathsf{y}_h - \mathbf{\Psi}_\ell \mathsf{y}_\ell)_i\big| \le \frac{C_1\|\mathrm{Res}_\ell^y(\mu)\|_{V_h'} + \varepsilon_{\tilde{\mathsf{M}}}(y_h - y_\ell)}{\tilde{\mathsf{M}}_{ii}^{1/2}}.$$

Accordingly, we obtain that

$$\begin{aligned}(\mathsf{y}_h)_i &\le (\mathbf{\Psi}_\ell \mathsf{y}_\ell)_i + \big|(\mathsf{y}_h - \mathbf{\Psi}_\ell \mathsf{y}_\ell)_i\big| \le (\mathbf{\Psi}_\ell \mathsf{y}_\ell)_i + \frac{C_1\|\mathrm{Res}_\ell^y(\mu)\|_{V_h'} + \varepsilon_{\tilde{\mathsf{M}}}(y_h - y_\ell)}{\tilde{\mathsf{M}}_{ii}^{1/2}}, \\ (\mathsf{y}_h)_i &\ge (\mathbf{\Psi}_\ell \mathsf{y}_\ell)_i - \big|(\mathsf{y}_h - \mathbf{\Psi}_\ell \mathsf{y}_\ell)_i\big| \ge (\mathbf{\Psi}_\ell \mathsf{y}_\ell)_i - \frac{C_1\|\mathrm{Res}_\ell^y(\mu)\|_{V_h'} + \varepsilon_{\tilde{\mathsf{M}}}(y_h - y_\ell)}{\tilde{\mathsf{M}}_{ii}^{1/2}}.\end{aligned}$$

With $\mathsf{g}_{\varepsilon_{\tilde{\mathsf{M}}}}^{\pm}$ introduced in (1.15) and the splitting of the nonpositive and positive components of the RB solution, we combine these estimates to finalize the proof be-

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

cause

$$\begin{aligned}
&((\Theta(\boldsymbol{\Psi}_\ell \mathsf{y}_\ell) - \Theta(\mathsf{y}_h))\boldsymbol{\Psi}_\ell \mathsf{p}_\ell)^T \tilde{\mathsf{M}}((\Theta(\boldsymbol{\Psi}_\ell \mathsf{y}_\ell) - \Theta(\mathsf{y}_h))\boldsymbol{\Psi}_\ell \mathsf{p}_\ell) \\
&\quad \leq ((\Theta(\mathsf{g}^+_{\varepsilon_{\tilde{\mathsf{M}}}}))\boldsymbol{\Psi}_\ell \mathsf{p}_\ell)^T \tilde{\mathsf{M}}(\Theta_1(-\boldsymbol{\Psi}_\ell \mathsf{y}_\ell))((\Theta(\mathsf{g}^+_{\varepsilon_{\tilde{\mathsf{M}}}})))\boldsymbol{\Psi}_\ell \mathsf{p}_\ell) \\
&\qquad + ((\Theta(\boldsymbol{\Psi}_\ell \mathsf{y}_\ell) - \Theta(\mathsf{g}^-_{\varepsilon_{\tilde{\mathsf{M}}}}))\boldsymbol{\Psi}_\ell \mathsf{p}_\ell)^T \tilde{\mathsf{M}}(\Theta(\boldsymbol{\Psi}_\ell \mathsf{y}_\ell))((\Theta(\boldsymbol{\Psi}_\ell \mathsf{y}_\ell) - \Theta(\mathsf{g}^-_{\varepsilon_{\tilde{\mathsf{M}}}}))\boldsymbol{\Psi}_\ell \mathsf{p}_\ell). \qquad \square
\end{aligned}$$

The term $\Delta^p_{w,\varepsilon_{\tilde{\mathsf{M}}}}(V_\ell, \mu)$ still depends on FE solutions, which are costly to compute and, to make matters worse, enter in the complicated grid-dependent quadrature error. Its evaluation therefore remains intractable in practice, unless a tight upper bound on the quadrature error can be derived. We do not provide such a result and do not expect that it is easily obtainable. In the numerical results of the next subsection, we instead assume that the mass lumping quadrature error is small compared to the other error sources in our computation and use the term

$$\begin{aligned}
\Delta^p_w(V_\ell, \mu) := C_1(&\gamma \Delta^y_w(V_\ell, \mu) + \|\mathrm{Res}^p_\ell(\mu)\|_{V_h'} \\
&+ [(\Theta(\mathsf{g}^+)\boldsymbol{\Psi}_\ell \mathsf{p}_\ell)^T \tilde{\mathsf{M}} \Theta_1(-\boldsymbol{\Psi}_\ell \mathsf{y}_\ell)(\Theta(\mathsf{g}^+)\boldsymbol{\Psi}_\ell \mathsf{p}_\ell) \\
&+ ((\Theta(\boldsymbol{\Psi}_\ell \mathsf{y}_\ell) - \Theta(\mathsf{g}^-))\boldsymbol{\Psi}_\ell \mathsf{p}_\ell)^T \tilde{\mathsf{M}} \Theta(\boldsymbol{\Psi}_\ell \mathsf{y}_\ell)((\Theta(\boldsymbol{\Psi}_\ell \mathsf{y}_\ell) - \Theta(\mathsf{g}^-))\boldsymbol{\Psi}_\ell \mathsf{p}_\ell)]^{\frac{1}{2}})
\end{aligned}$$

as an adjoint error indicator, which is the result of assuming that all quadrature errors in $\Delta^p_{w,\varepsilon_{\tilde{\mathsf{M}}}}$ vanish and which is easily evaluable. Interestingly, we obtain the reproduction of the solution results analogously to Corollary 1.2.6 in the PDE analysis for both combined error indicators $\Delta^{y,p}_{w,\varepsilon_{\tilde{\mathsf{M}}}}(V_\ell, \mu) = \Delta^y_w(V_\ell, \mu) + \Delta^p_{w,\varepsilon_{\tilde{\mathsf{M}}}}(V_\ell, \mu)$ and $\Delta^{y,p}_w(V_\ell, \mu) = \Delta^y_w(V_\ell, \mu) + \Delta^p_w(V_\ell, \mu)$, which motivates us to use the latter in the numerics.

**Corollary 1.3.5.** *Let the assumptions of Lemma* 1.3.4 *hold, and let* $\mu \in \mathbb{R}^p$ *with* $y_h, p_h \in V_\ell$. *Then*

1) $y_h = y_\ell$ *and* $p_h = p_\ell$,
2) $\Delta^{y,p}_{w,\varepsilon_{\tilde{\mathsf{M}}}}(V_\ell, \mu) = \Delta^{y,p}_w(V_\ell, \mu) = 0$.

*Proof.*

1) $y_h = y_\ell$ follows from Corollary 1.2.6-1). For the adjoint, we use the fact that $y_h = y_\ell$ and $e_p = p_h - p_\ell \in V_\ell$ to conclude that

$$0 \leq \langle \nabla e_p, \nabla e_p \rangle_{L^2} + \langle \mathbb{1}_{\{y_\ell > 0\}} e_p, e_p \rangle_{L^2} = j'(y_h) - j'(y_\ell) = j'(y_\ell) - j'(y_\ell) = 0,$$

   where the first inequality follows from the coercivity of the adjoint equation.
2) The equality $\Delta^y_w(V_\ell, \mu) = 0$ follows from Corollary 1.2.6-2). We infer from 1) that since the quadrature error of the constant zero function vanishes, $\Delta^{y,p}_{w,\varepsilon_{\tilde{\mathsf{M}}}}(V_\ell, \mu) = \Delta^{y,p}_w(V_\ell, \mu)$, and additionally $\|\mathrm{Res}^p_\ell(\mu)\|_{V_h'} = 0$. Furthermore, since $\Delta^y_w(V_\ell, \mu) = 0$, we can conclude that $\mathsf{g}^+ = \mathsf{g}^- = \boldsymbol{\Psi}_\ell \mathsf{y}_\ell$. This implies $\Delta^p_w(V_\ell, \mu) = \Delta^p_{w,\varepsilon_{\tilde{\mathsf{M}}}}(V_\ell, \mu) = 0$, and thus $\Delta^{y,p}_{w,\varepsilon_{\tilde{\mathsf{M}}}}(V_\ell, \mu) = \Delta^{y,p}_w(V_\ell, \mu) = 0$. $\square$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

### 1.3.4 Numerical results for the optimal control problem

Compared to the analytical results for the model order reduction of the PDE constraint in Section 1.2, there are quite few analytical results on the optimal control problem in the previous subsections. This is largely due to the fact that the indicator function in the adjoint equation, in principle, does not allow a direct derivation of an analytical a posteriori error estimate. However, with the additional use of the discretization structure, the error indicator derived in Lemma 1.3.4 offers a way to deal with error indication in the numerics. Furthermore, the parameter space is no longer necessarily compact in the optimization setting, which is typically a fundamental assumption for the offline/online reduced basis method. The adaptive RB approach introduced in Section 1.3.2 is designed to circumvent this issue. In this subsection, we present numerical results based on the true $V$-error computation and the error indicator derived in Lemma 1.3.4, comparing the two approaches for a truly finite-dimensional parameter space and showing a promising performance of the adaptive RB method for an inherently infinite-dimensional control problem, which usually breaks offline/online RB approaches.

For the remainder of this section, we assume that $j(y) := \frac{1}{2}\|y - y^d\|_{L^2}^2$ and write $\mathsf{B} \in \mathbb{R}^{N\times p}$ for the matrix corresponding to the discretized linear operator $\mathcal{B}\colon \mathbb{R}^p \to V'$. As in Section 1.2.2, we fix $\Omega = (0,1)^2$ and consider $P_1$-type finite elements on a Friedrichs–Keller triangulation of the domain in the discretization. We maintain the mass lumping procedure in the nonlinear max term and indicator function and from the discretization procedures applied to (1.13) obtain the following FE and RB systems:

$$\begin{aligned}
\mathsf{K}\bar{\mathsf{y}}_h + \tilde{\mathsf{M}}\max\{0,\bar{\mathsf{y}}_h\} &= \mathsf{B}\bar{\mu} \quad \text{in } \mathbb{R}^N,\\
\mathsf{K}\bar{\mathsf{p}}_h + \tilde{\mathsf{M}}\Theta(\bar{\mathsf{y}}_h)\bar{\mathsf{p}}_h &= \mathsf{M}(\bar{\mathsf{y}}_h - \mathsf{y}^d) \quad \text{in } \mathbb{R}^N,\\
\mathsf{B}^T\bar{\mathsf{p}}_h + \sigma\mathsf{A}\bar{\mu} &= 0 \quad \text{in } \mathbb{R}^p,
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\Psi}_\ell^T\mathsf{K}\boldsymbol{\Psi}_\ell\bar{\mathsf{y}}_\ell + \boldsymbol{\Psi}_\ell^T\tilde{\mathsf{M}}\max\{0,\boldsymbol{\Psi}_\ell\bar{\mathsf{y}}_\ell\} &= \boldsymbol{\Psi}_\ell^T\mathsf{B}\bar{\mu} \quad \text{in } \mathbb{R}^\ell,\\
\boldsymbol{\Psi}_\ell^T\mathsf{K}\boldsymbol{\Psi}_\ell\bar{\mathsf{p}}_\ell + \boldsymbol{\Psi}_\ell^T\tilde{\mathsf{M}}\Theta(\boldsymbol{\Psi}_\ell\bar{\mathsf{y}}_\ell)\bar{\mathsf{p}}_\ell &= \boldsymbol{\Psi}_\ell^T\mathsf{M}(\boldsymbol{\Psi}_\ell\bar{\mathsf{y}}_\ell - \mathsf{y}^d) \quad \text{in } \mathbb{R}^\ell,\\
\mathsf{B}^T\boldsymbol{\Psi}_\ell\bar{\mathsf{p}}_\ell + \sigma\mathsf{A}\bar{\mu} &= 0 \quad \text{in } \mathbb{R}^p.
\end{aligned}$$

These finite-dimensional systems are solved with the PSN method described in the beginning of this section. The FE and RB system matrices at iterates $(y_h, p_h, \mu)$, $(y_\ell, p_\ell, \mu)$ read as

$$\begin{pmatrix}
\mathsf{K} + \tilde{\mathsf{M}}\Theta(\mathsf{y}_h) & 0 & -\mathsf{B}\\
-\mathsf{M} & \mathsf{K} + \tilde{\mathsf{M}}\Theta(\mathsf{y}_h) & 0\\
0 & \mathsf{B}^T & \sigma\mathsf{A}
\end{pmatrix},$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$$\begin{pmatrix} \Psi_\ell^T \mathsf{K} \Psi_\ell + \Psi_\ell^T \tilde{\mathsf{M}} \Theta(\Psi_\ell \mathsf{y}_\ell) \Psi_\ell & 0 & -\Psi_\ell^T \mathsf{B} \\ -\Psi_\ell^T \mathsf{M} \Psi_\ell & \Psi_\ell^T \mathsf{K} \Psi_\ell + \Psi_\ell^T \tilde{\mathsf{M}} \Theta(\Psi_\ell \mathsf{y}_\ell) \Psi_\ell & 0 \\ 0 & \mathsf{B}^T \Psi_\ell & \sigma \mathsf{A} \end{pmatrix}.$$

As for the numerical results in Section 1.2, our code is implemented in Python3 and uses FENICS [1] for the matrix assembly. Sparse memory management and computations are implemented with SciPy [24]. All computations below were run on an Ubuntu 18.04 notebook with 12 GB main memory and an Intel Core i7-4600U CPU.

**Example 1.3.1.** For the first numerical example of this section, we reuse the setting of Example 1.2.1, i. e., we set $\mathcal{P} = [-2, 2]^2$, $c(\mu) = 1$, $a(\mu) = 10$ (note that this does not change the analysis) and the right-hand side as defined in (1.11). Additionally, we use $y^d(\boldsymbol{x}) = (\frac{1}{2} - x_1)\sin(\pi x_1)\sin(\pi x_2)$ for $\boldsymbol{x} = (x_1, x_2) \in \Omega$, $\mathsf{A} = \mathsf{I}_2$ the identity in $\mathbb{R}^2$, and $\sigma = 10^{-3}$. As previously, we fix a training set of 121 and a test set of 196 equidistant points in $\mathcal{P}$, where both sets are chosen disjoint. The initial value in all computations is set to $\mu_0 = [1, 1]$. We keep the tolerances fixed at $10^{-4}$ for the RB error, at $10^{-8}$ for the semismooth Newton method, and now additionally at $10^{-6}$ for the PSN method. The desired state, solution state, and solution adjoint computed with finite elements are shown in Figure 1.4. The results of the different approaches for different grids with step sizes $h = 1/100, 1/200, 1/400$ per dimension can be found in Table 1.5. For the adaptive RB approach, all parameter combinations $(\eta, n_{\text{fix}}) \in \{0.5,\ 0.25\} \times \{1, 2, 3, 4\}$ are tested, and the best, worst, and average performances are stated. We can again note that all the $V$-errors are below the given tolerance of the offline/online and adaptive RB method, respectively. The error indicator's overestimation leads to only slightly smaller errors in the offline/online approach but produce significantly lower errors, larger reduced bases, and slower speedup in the adaptive RB approach, compared to the true error computation. Additionally, the reduced bases for the adaptive RB approach are much smaller than those for the offline/online approach and contain only six or eight elements, depending on the utilized error indicator. This also suggests that a basis of only two more elements is enough to largely reduce the $V$-errors in the adaptive case. Furthermore, the size of the basis in the adaptive RB seems to be independent of the chosen parameters $\eta$ and $n_{\text{fix}}$ and the step size $h$ in the adaptive RB approach. As before, the number of PSN iterations for solving the FE/RB systems appears

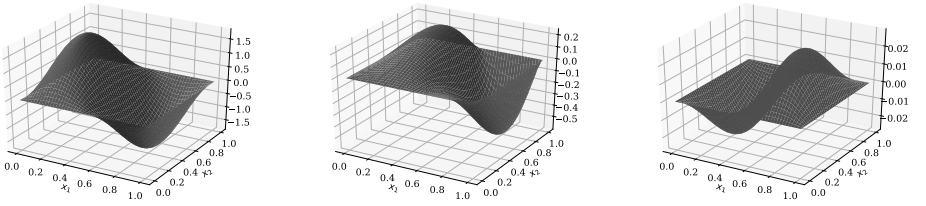


**Figure 1.4:** Example 1.3.1. Desired state (left), FE solution state (middle), and FE solution adjoint (right) for $1/h = 400$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Table 1.5:** Example 1.3.1. The first part of the table shows the number of required pseudo-semismooth Newton iterations for computing the FE solution to the first-order system and the required time. The second and third parts show the performance of the offline/online RB approach including the offline computation time $t_{\text{off}}$ The fourth and fifth parts show the best, worst, and average performance of the adaptive RB algorithm over all $(\eta, n_{\text{fix}}) \in \{0.5,\ 0.25\} \times \{1, 2, 3, 4\}$ using the true $V$-error and error estimator, respectively.

| **FE** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **1/h** | | **it.** | **time (s)** | | | | |
| 100 | | 4 | 8.96 | | | | |
| 200 | | 4 | 146.28 | | | | |
| 400 | | 4 | 4008.98 | | | | |
| **offline/online RB (true $V$-error)** | | | | | | | |
| **1/h** | | **it.** | **speedup** | $t_{\text{off}}$ (s) | $\|e_y\|_V$ | $\|e_p\|_V$ | $\|\Psi\|$ |
| 100 | | 4 | 70.67 | 98.03 | $2.67 \cdot 10^{-5}$ | $3.75 \cdot 10^{-5}$ | 60 |
| 200 | | 4 | 373.75 | 523.89 | $1.28 \cdot 10^{-5}$ | $3.42 \cdot 10^{-5}$ | 58 |
| 400 | | 4 | 1544.93 | 3427.74 | $7.66 \cdot 10^{-6}$ | $3.36 \cdot 10^{-5}$ | 58 |
| **offline/online RB (indicator)** | | | | | | | |
| **1/h** | | **it.** | **speedup** | $t_{\text{off}}$ (s) | $\|e_y\|_V$ | $\|e_p\|_V$ | $\|\Psi\|$ |
| 100 | | 4 | 30.45 | 1054.56 | $1.28 \cdot 10^{-5}$ | $2.25 \cdot 10^{-5}$ | 84 |
| 200 | | 4 | 235.92 | 6770.66 | $4.75 \cdot 10^{-6}$ | $2.09 \cdot 10^{-5}$ | 84 |
| 400 | | 4 | 1105.49 | 47892.43 | $2.09 \cdot 10^{-6}$ | $1.98 \cdot 10^{-5}$ | 84 |
| **adaptive RB (true $V$-error)** | | | | | | | |
| **1/h** | $(\eta, n_{\text{fix}})$ | **it.** | **speedup** | | $\|e_y\|_V$ | $\|e_p\|_V$ | $\|\Psi\|$ |
| | best (0.25, 3) | 9 | 3.81 | | $2.38 \cdot 10^{-5}$ | $5.69 \cdot 10^{-6}$ | 6 |
| 100 | worst (0.25, 1) | 6 | 2.14 | | $2.08 \cdot 10^{-5}$ | $5.63 \cdot 10^{-6}$ | 6 |
| | avg. | 8.25 | 3.35 | | $2.22 \cdot 10^{-5}$ | $5.65 \cdot 10^{-6}$ | 6 |
| | best (0.25, 2) | 6 | 10.35 | | $1.96 \cdot 10^{-5}$ | $4.44 \cdot 10^{-6}$ | 6 |
| 200 | worst (0.5, 1) | 6 | 5.88 | | $1.96 \cdot 10^{-5}$ | $4.44 \cdot 10^{-6}$ | 6 |
| | avg. | 8.25 | 9.18 | | $2.12 \cdot 10^{-5}$ | $4.64 \cdot 10^{-6}$ | 6 |
| | best (0.25, 3) | 9 | 40.87 | | $2.26 \cdot 10^{-5}$ | $4.23 \cdot 10^{-6}$ | 6 |
| 400 | worst (0.5, 1) | 6 | 22.85 | | $1.94 \cdot 10^{-5}$ | $3.83 \cdot 10^{-6}$ | 6 |
| | avg. | 8.25 | 36.14 | | $2.09 \cdot 10^{-5}$ | $3.90 \cdot 10^{-6}$ | 6 |
| **adaptive RB (indicator)** | | | | | | | |
| **1/h** | $(\eta, n_{\text{fix}})$ | **it.** | **speedup** | | $\|e_y\|_V$ | $\|e_p\|_V$ | $\|\Psi\|$ |
| | best (0.5, 2) | 12 | 2.73 | | $3.87 \cdot 10^{-10}$ | $7.35 \cdot 10^{-10}$ | 8 |
| 100 | worst (0.5, 1) | 6 | 2.35 | | $1.09 \cdot 10^{-8}$ | $6.05 \cdot 10^{-8}$ | 8 |
| | avg. | 10.25 | 2.63 | | $3.02 \cdot 10^{-9}$ | $1.64 \cdot 10^{-8}$ | 8 |
| | best (0.25, 2) | 8 | 7.22 | | $7.14 \cdot 10^{-11}$ | $4.38 \cdot 10^{-10}$ | 8 |
| 200 | worst (0.5, 1) | 6 | 6.26 | | $8.73 \cdot 10^{-9}$ | $4.86 \cdot 10^{-8}$ | 8 |
| | avg. | 10.25 | 7.07 | | $1.35 \cdot 10^{-9}$ | $7.81 \cdot 10^{-9}$ | 8 |
| | best (0.25, 2) | 8 | 28.14 | | $2.95 \cdot 10^{-11}$ | $4.13 \cdot 10^{-10}$ | 8 |
| 400 | worst (0.5, 1) | 6 | 24.38 | | $2.38 \cdot 10^{-9}$ | $2.24 \cdot 10^{-6}$ | 8 |
| | avg. | 10.25 | 27.50 | | $4.12 \cdot 10^{-10}$ | $4.74 \cdot 10^{-9}$ | 8 |

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

to be independent of the step size $h$ and is the same for the full FE system and the offline/online RB approach. Of course, this number differs for the adaptive RB approach with its prescribed number of $n_{\text{fix}}$ fixed PSN steps per basis update. In general, the speedup in the online phase disregarding the offline computational time in the classical RB approach is, as expected, far greater than that of the adaptive RB approach (up to two orders of magnitude). Also, as in the previous examples, the true $V$-error outperforms the error indicator in the RB greedy procedure. It is worth mentioning that for step size $h = 1/400$, the offline/online RB approach with true $V$-error still requires less total computation time than the FE approach even if the offline time is taken into account. For the adaptive RB approach, the results show significant speedups in all cases, which increase with a factor of about three to four when the step size $h$ is divided by two. When we consider the best and average performance, computations using the true $V$-error are always faster compared to the error indicator, but when we consider the worst performance, the error indicator always slightly outperforms the true $V$-error. Note that the average speedup is much closer to the best speedup than it is to the worst, especially for the true error. Additionally, the worst performance is always achieved for the case $n_{\text{fix}} = 1$ and mostly for $\eta = 0.5$, which can be expected, since in this case an evaluation of the (costly) error indicator is necessary in every iteration. Accordingly, choices $n_{\text{fix}} \in \{2, 3\}$, where the algorithm reduces the norm of the residual in the new updated RB space considerably before updating the basis further, are arguably more practical.

We conclude that the adaptive RB approach can be suitable to solve the first-order system and outperform the classical offline/online approach, which is generally a suboptimal choice when only few solves of the problem are required. The new method computes solutions whose approximation quality of the FE solutions can be controlled well at a speedup of up to 40 using a reduced basis with comparatively few elements.

The performance and effectivity for the error indicator are shown in Table 1.6. We can observe that here the effectivity appears to be mesh dependent, in contrast to the effectivity of the analytical estimator in Section 1.2. This is to be expected because the indicator includes grid-dependent parts and the mass lumping error is ignored. Regardless, even for large $h$, the performance is reasonable, and it is improved for finer grids. The lower speedup of approximately two, compared to the speedup of four in Section 1.2, is due to the fact that the adjoint equation is linear. Therefore solving the

**Table 1.6:** Example 1.3.1. Performance of true error and error indicator with reduced basis from strong greedy on the test set.

| 1/h | avg. true $V$-error | avg. effectivity | time true $V$-error (s) | speed-up indicator |
|---|---|---|---|---|
| 100 | $4.85 \cdot 10^{-5}$ | 36.85 | 114.36 | 2.32 |
| 200 | $4.58 \cdot 10^{-5}$ | 70.08 | 697.43 | 2.27 |
| 400 | $4.53 \cdot 10^{-5}$ | 133.18 | 4933.28 | 2.18 |

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

linear system for the error estimator does not save time compared to solving the adjoint equation. Of course, the computation of the state remains nonlinear, and hence the remaining speedup.

**Example 1.3.2.** The second example in this section is designed to show that the adaptive approach presented above can overcome the curse of dimensionality that comes with classical offline/online approaches for large parameter spaces. We consider an FE discretized approximation of the standard optimal control problem where (**P**) is considered with distributed $L^2(\Omega)$ controls on the right-hand side of the PDE and an $\|\cdot\|_{L^2}$-regularization term in the cost functional – and therefore an inherently infinite-dimensional setting. We set $p = N$, $\mathsf{A} = \mathsf{B} = \mathsf{M}$ as the mass matrix, and $y^d(x_1, x_2)$ as the solution of (1.12) in Section 1.2.2 for $\mu = 0$, i. e., the function consisting of three bumps on three quarters of the domain. This implies that

$$\bar{p}_h = \frac{1}{\sigma}\bar{\mu}$$

in the FE system. Accordingly, the reduced basis can be used to approximate the parameter space itself because the adjoint and parameter are linearly dependent in the solution. Additionally, we set $\sigma = 10^{-4}$, and for increased influence of the nonsmooth term, we set $a(\mu) = 100$ as the scaling parameter in front of the max term (note that this does not change the analysis). The different settings for $\eta$, $n_{\text{fix}}$, and the grids remain the same as they were in Example 1.3.1. In Figure 1.5, the desired state, FE solution state, and FE solution adjoint are shown.

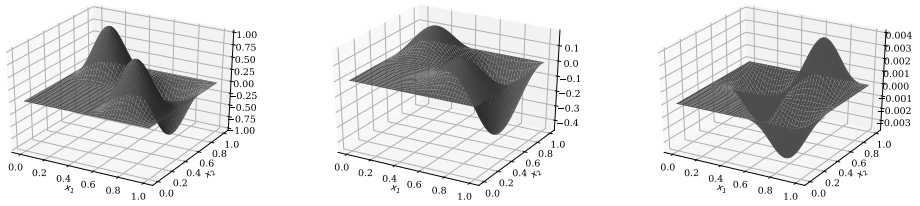



**Figure 1.5:** Example 1.3.2. Desired state (left), FE solution state (middle), and FE solution adjoint (right) for $1/h = 400$.

**Remark 1.3.6.** Note that although $\mathsf{A} = \mathsf{B} = \mathsf{M}$ will not be a diagonal matrix as asked for in Assumption 1.1.1, the diagonality requirement has been solely for the purpose of obtaining a reasonable sense of stationarity for the solutions of (1.13) in the finite-dimensional control case. In the case of $L^2(\Omega)$ controls with $L^2(\Omega)$ regularization treated in this example, (1.13) is a strong stationarity system by the results shown in [9].

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

For the reasons stated above, the offline/online RB approach will not be compared here. Instead, only the adaptive RB approach and the standard FE solution process are compared in Table 1.7.

**Table 1.7:** Example 1.3.2. The first part of the table shows the number of required PSN iterations for computing the FE solution to the first-order system and the required time. The second and third parts show the best, worst, and average performance of the adaptive RB algorithm over all $(\eta, n_{\text{fix}}) \in \{0.5,\ 0.25\} \times \{1, 2, 3, 4\}$ using the true $V$-error and error indicator, respectively.

| **FE** | | | | | | |
|---|---|---|---|---|---|---|
| **1/h** | | **it.** | **time (s)** | | | |
| 100 | | 5 | 10.80 | | | |
| 200 | | 5 | 84.18 | | | |
| 400 | | 5 | 765.69 | | | |
| **adaptive RB (true $V$-error)** | | | | | | |
| **1/h** | **$(\eta, n_{\text{fix}})$** | **it.** | **speedup** | **$\|e_y\|_V$** | **$\|e_p\|_V$** | **$\|\Psi\|$** |
| | best (0.5, 3) | 18 | 2.63 | $2.28 \cdot 10^{-5}$ | $9.27 \cdot 10^{-8}$ | 12 |
| 100 | worst (0.25, 1) | 9 | 1.47 | $3.27 \cdot 10^{-5}$ | $4.79 \cdot 10^{-7}$ | 12 |
| | avg. | 16 | 2.14 | $2.22 \cdot 10^{-5}$ | $1.37 \cdot 10^{-7}$ | 12 |
| | best (0.25, 3) | 18 | 2.90 | $1.41 \cdot 10^{-5}$ | $8.47 \cdot 10^{-7}$ | 12 |
| 200 | worst (0.5, 1) | 9 | 1.73 | $1.74 \cdot 10^{-5}$ | $8.70 \cdot 10^{-7}$ | 12 |
| | avg. | 15.87 | 2.43 | $1.35 \cdot 10^{-5}$ | $5.01 \cdot 10^{-7}$ | 12 |
| | best (0.25, 4) | 24 | 3.92 | $1.02 \cdot 10^{-5}$ | $5.24 \cdot 10^{-8}$ | 12 |
| 400 | worst (0.5, 1) | 9 | 2.32 | $9.79 \cdot 10^{-6}$ | $6.47 \cdot 10^{-8}$ | 12 |
| | avg. | 15.87 | 3.31 | $1.12 \cdot 10^{-5}$ | $1.13 \cdot 10^{-7}$ | 12 |
| **adaptive RB (indicator)** | | | | | | |
| **1/h** | **$(\eta, n_{\text{fix}})$** | **it.** | **speedup** | **$\|e_y\|_V$** | **$\|e_p\|_V$** | **$\|\Psi\|$** |
| | best (0.5, 3) | 21 | 1.88 | $4.11 \cdot 10^{-7}$ | $3.05 \cdot 10^{-9}$ | 14 |
| 100 | worst (0.5, 1) | 10 | 1.51 | $1.58 \cdot 10^{-6}$ | $7.14 \cdot 10^{-9}$ | 16 |
| | avg. | 18.12 | 1.72 | $1.17 \cdot 10^{-6}$ | $7.84 \cdot 10^{-9}$ | 14.25 |
| | best (0.2, 3) | 21 | 2.06 | $1.01 \cdot 10^{-5}$ | $8.55 \cdot 10^{-7}$ | 14 |
| 200 | worst (0.5, 1) | 9 | 1.88 | $1.09 \cdot 10^{-5}$ | $8.55 \cdot 10^{-7}$ | 14 |
| | avg. | 18 | 1.99 | $9.62 \cdot 10^{-6}$ | $8.56 \cdot 10^{-7}$ | 14 |
| | best (0.25, 3) | 21 | 2.76 | $2.57 \cdot 10^{-7}$ | $1.17 \cdot 10^{-9}$ | 14 |
| 400 | worst (0.5, 1) | 10 | 2.27 | $1.41 \cdot 10^{-6}$ | $6.47 \cdot 10^{-9}$ | 16 |
| | avg. | 18.12 | 2.64 | $4.98 \cdot 10^{-6}$ | $4.56 \cdot 10^{-8}$ | 14.25 |

First, note that again the number of PSN iterations seems to be mesh-independent in the FE formulations. The iterations in the adaptive approach also show the mesh independence when best performances are compared with best performances of finer/coarser grids, etc. This time, though, the number of (average) iterations for the adaptive RB approach is about twice as large as in the previous example. This is also the case for the size of the reduced basis, which always contains 12 elements for the true error and 14–16 elements for the error indicator. Compared to the previ-

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

ous example, the additional basis elements do not increase the quality of the results significantly. A richer reduced basis was to be expected due to the underlying high-dimensional structure of the problem. The resulting speedups are not as high as in the previous example and do not increase as quickly as before when $h$ is decreased. Whereas it is reasonable to assume that this is due to the difficulties with the reduced basis approximation, note that the computational time for solving the FE systems does not increase as strongly as before as $h$ is decreased either – compare 765.69 seconds with 4008.98 seconds total in the last example for the finest grid. Nonetheless, a speedup of close to four can be obtained, and the FE setting never outperforms the adaptive RB approach. This shows that the adaptive RB approach can be used as a reasonable approach when RB model order reduction on very high-dimensional parameter spaces is desired.

## 1.4 Conclusion

We have established an a posteriori error estimator and presented corresponding numerical considerations for classical offline/online RB approaches for efficiently solving a generalized version of the constraining nonsmooth PDE of (**P**). The results show promising speedup, suggesting that the nonsmooth and nonlinear behavior of the max term in the PDE is in fact the part that is most difficult to capture in the RB approximation. For the first-order system to the optimal control problem (**P**) itself, we have proposed a novel adaptive RB pseudo-semismooth Newton approach, which creates the required reduced bases adaptively as the PSN algorithm progresses, using local information along the PSN iterates. The adaptive approach is complementary to standard offline/online approaches in the sense that it offers reasonable to large speedup in online computations without an additional penalty of offline computation times.

## Bibliography

[1] M. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, and G. N. Wells. The FEniCS project version 1.5. *Arch. Numer. Softw.*, 3(100):9–23, 2015.

[2] M. Barrault, Y. Maday, N. C. Nguyen, and A. T Patera. An empirical interpolation method: application to efficient reduced-basis discretization of partial differential equations. *C. R. Math.*, 339(9):667–672, 2004.

[3] P. Benner, M. Ohlberger, A. Cohen, and K. Willcox. *Model Reduction and Approximation: Theory and Algorithms*. Computational Science and Engineering. SIAM, 2017.

[4] M. Bernreuther. RB-based PDE-constrained non-smooth optimization. Master's thesis, Universität Konstanz, 2019.

[5] C. Canuto, T. Tonn, and K. Urban. A posteriori error analysis of the reduced basis method for nonaffine parametrized nonlinear PDEs. *SIAM J. Numer. Anal.*, 47(3):2001–2022, 2009.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[6] S. Chaturantabut and D. C. Sorensen. Nonlinear model reduction via discrete empirical interpolation. *SIAM J. Sci. Comput.*, 32(5):2737–2764, 2010.
[7] S. Chaturantabut and D. C. Sorensen. A state space estimate for POD-DEIM nonlinear model reduction. *SIAM J. Numer. Anal.*, 50(1):46–63, 2012.
[8] C. Christof, C. Clason, C. Meyer, and S. Walther. Optimal control of a non-smooth semilinear elliptic equation. *Math. Control Relat. Fields*, 8:247–276, 2018.
[9] C. Christof and G. Müller. Multiobjective optimal control of a non-smooth semilinear elliptic partial differential equation. Technical report, 2020. Submitted.
[10] D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Classics in Mathematics. Springer, 2001.
[11] M. A. Grepl, Y. Maday, N. C. Nguyen, and A. T. Patera. Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations. *ESAIM: Math. Model. Numer. Anal.*, 41(3):575–605, 2007.
[12] B. Haasdonk. A tutorial on RB-methods. In P. Benner, A. Cohen, M. Ohlberger, and K. Willcox, editors, *Model Reduction and Approximation: Theory and Algorithms*. Computational Science & Engineering, pages 67–138. SIAM, 2017.
[13] J. Hesthaven, G. Rozza, and B. Stamm. *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*. SpringerBriefs in Mathematics. Springer, 2016.
[14] M. Hinze and D. Korolev. Reduced basis methods for quasilinear elliptic PDEs with applications to permanent magnet synchronous motors. Technical report, 2020. Submitted.
[15] M. Kärcher, Z. Tokoutsi, M. Grepl, and K. Veroy. Certified reduced basis methods for parametrized elliptic optimal control problems with distributed controls. *J. Sci. Comput.*, 75:276–307, 2018.
[16] F. Kikuchi, K. Nakazato, and T. Ushijima. Finite element approximation of a nonlinear eigenvalue problem related to MHD equilibria. *Jpn. J. Appl. Math.*, 1(2):369–403, 1984.
[17] F. Negri, G. Rozza, A. Manzoni, and A. Quarteroni. Reduced basis method for parametrized elliptic optimal control problems. *SIAM J. Sci. Comput.*, 35:A2316–A2340, 2013.
[18] A. T. Patera and G. Rozza. *Reduced Basis Approximation and A Posteriori Error Estimation for Parametrized Partial Differential Equations*. MIT Pappalardo Graduate Monographs in Mechanical Engineering, 2007.
[19] E. Qian, M. Grepl, K. Veroy, and K. Willcox. A certified trust region reduced basis approach to PDE-constrained optimization. *SIAM J. Sci. Comput.*, 39(5):S434–S460, 2017.
[20] A. Quarteroni, A. Manzoni, and F. Negri. *Reduced Basis Methods for Partial Differential Equations*. Unitext Series. Springer, 2016.
[21] J. Rappaz. Approximation of a nondifferentiable nonlinear problem related to MHD equilibria. *Numer. Math.*, 45(1):117–133, 1984.
[22] R. Temam. A non-linear eigenvalue problem: the shape at equilibrium of a confined plasma. *Arch. Ration. Mech. Anal.*, 60(1):51–73, 1976.
[23] K. Veroy, C. Prud'homme, D. V. Rovas, and A. T. Patera. A posteriori error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations. In *16th AIAA Computational Fluid Dynamics Conference*, 2003, Orlando, United States, 2003.
[24] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, 17:261–272, 2020.
[25] J. Xin. *An Introduction to Fronts in Random Media*, volume 5 of Surveys and Tutorials in the Applied Mathematical Sciences. Springer, 2009.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Arthur Bottois

# 2 Pointwise moving control for the 1-D wave equation

## Numerical approximation and optimization of the support

**Abstract:** We consider the exact null controllability of the 1-D wave equation with an interior pointwise control acting on a moving point $(\gamma(t))_{t\in[0,T]}$. We approximate a control of minimal norm through a mixed formulation solved by using a conformal space-time finite element method. We then introduce a gradient-type approach to optimize the trajectory $\gamma$ of the control point. Several experiments are discussed.

**Keywords:** exact controllability, wave equation, pointwise control, mixed formulation, finite element approximation

**MSC 2010:** 49Q10, 93C20

## 2.1 Introduction

Let $T > 0$. We consider the linear one-dimensional wave equation in the interval $\Omega = (0,1)$ with a pointwise control $v$ acting on a moving point $x = \gamma(t)$, $t \in [0,T]$. The state equation reads

$$\begin{cases} y_{tt} - y_{xx} = v(t)\delta_{\gamma(t)}(x) & \text{in } Q_T = \Omega \times (0,T), \\ y = 0 & \text{on } \Sigma_T = \partial\Omega \times (0,T), \\ (y, y_t)(\cdot, 0) = (y_0, y_1) & \text{in } \Omega. \end{cases} \tag{2.1}$$

Here, $\delta_{\gamma(t)}$ is the Dirac measure at $x = \gamma(t)$, and $\gamma$ represents the trajectory in time of the control point. The curve $\gamma : [0,T] \to \Omega$ is assumed to be piecewise $C^1$. We also denote by $\mathbf{H}'$ the dual space of $\mathbf{H} := H^1(0,T)$. For $v \in \mathbf{H}'$, we refer to Section 2.2.1 for the well-posedness of (2.1).

The exact null controllability problem for (2.1) at time $T > 0$ is the following. *Given a trajectory $\gamma : [0,T] \to \Omega$, for any initial datum $(y_0, y_1) \in \mathbf{V} := L^2(\Omega) \times H^{-1}(\Omega)$, find a control $v \in \mathbf{H}'$ such that the corresponding solution $y$ of* (2.1) *satisfies*

$$(y, y_t)(\cdot, T) = (0,0) \quad \textit{in } \Omega.$$

**Arthur Bottois,** Université Clermont Auvergne, Laboratoire de Mathématiques Blaise Pascal, UMR 6620, 3 place Vasarely, 63178 Aubière, France, e-mail: arthur.bottois@uca.fr

https://doi.org/10.1515/9783110695984-002

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

As a consequence of the *Hilbert uniqueness method* (HUM) introduced by Lions [25], the controllability of (2.1) is equivalent to an observability inequality for the associated adjoint problem. Indeed, the state equation (2.1) is controllable if and only if there exists a constant $C_{\text{obs}}(\gamma) > 0$ such that

$$\|(\varphi_0, \varphi_1)\|^2_{\mathbf{W}} \leq C_{\text{obs}}(\gamma) \|\varphi(\gamma, \cdot)\|^2_{\mathbf{H}} \quad \forall (\varphi_0, \varphi_1) \in \mathbf{W} := H^1_0(\Omega) \times L^2(\Omega), \tag{2.2}$$

where $\varphi \in C([0,T]; H^1_0(\Omega)) \cap C^1([0,T]; L^2(\Omega))$ solves

$$L\varphi = 0 \text{ in } Q_T, \quad \varphi = 0 \text{ on } \Sigma_T, \quad (\varphi, \varphi_t)(\cdot, 0) = (\varphi_0, \varphi_1) \text{ in } \Omega. \tag{2.3}$$

Here the notation $\varphi(\gamma, \cdot)$ stands for the function $\varphi(\gamma(t), t)$, $t \in (0,T)$, whereas $L$ denotes the wave operator

$$L = \partial_t^2 - \partial_x^2.$$

Under additional assumptions on $\gamma$, a proof of (2.2) can be found in [8]. We emphasize that the observability constant $C_{\text{obs}}(\gamma)$ depends on the control trajectory $\gamma$. In what follows, we say that $\gamma$ is an *admissible trajectory* if the observability inequality (2.2) holds.

In this work, we investigate the issue of the numerical approximation of the control $\widehat{v}_\gamma$ of minimal $\mathbf{H}'$-norm and the associated controlled state. We also tackle the problem of optimizing the support of control, which is done numerically by minimizing the norm $\|\widehat{v}_\gamma\|_{\mathbf{H}'}$ with respect to the trajectory $\gamma$.

Let us now mention some references related to pointwise control.This problem arises naturally in practical situations where the size of the control domain is very small compared to the size of the physical system. For a stationary control point $\gamma \equiv x_0 \in \Omega$, the controllability of (2.1) depends strongly on the location of $x_0$ [14, 24, 26]. Indeed, we can show that controllability holds if and only if the controllability time $T$ is large enough, i. e., $T \geq 2|\Omega|$, and if there is no eigenfunction of the Dirichlet Laplacian vanishing at $x = x_0$. The constraint on $T$ is due to the finite speed of propagation of the solution of the wave equation (2.1). A point $x_0$ satisfying the previous spectral property is referred to as a *strategic point*. Furthermore, $x_0$ is a strategic point if and only if it is irrational with respect to the length of $\Omega$, making controllability very unstable. Consequently, controls acting on stationary points are usually difficult to implement in practice. It is often more convenient to control along curves for which the strategic point property holds a. e. in $[0,T]$.

For a moving control point $x = \gamma(t)$, several sufficient conditions to ensure controllability have been studied [1, 8, 20, 22]. In [22] the author proves the existence of controls in $L^2(0,T)$ acting on a point rapidly bouncing between two positions. In [8, Proposition 4.1] the author shows, using the d'Alembert formula, that the observability inequality (2.2) holds under some geometric restrictions on the trajectory $\gamma$. By duality this implies the existence of controls in $\mathbf{H}'$ for initial data in $\mathbf{V}$. The geometric

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

requirements are related to the usual *geometric control condition* (GCC) introduced for controls acting over domains $\omega \subset \Omega$ [3, 23]. Among the constraints given to guarantee that $\gamma$ is admissible, there must exist two constants $c_1, c_2 > 0$ and a finite number of subintervals $(I_j)_{0\le j\le J} \subset [0,T]$ such that, for each subinterval $I_j$, $\gamma \in C^1(I_j)$, $1 - |\gamma'|$ does not change sign in $I_j$ and $c_1 \le |\gamma'| \le c_2$ in $I_j$. The constants appearing in the proof of the observability inequality (2.2) depend only on $c_1$ and $c_2$ (see [8, Remark 4.2]). Thus it is possible to write a uniform observability inequality for trajectories in a suitable class, i. e., there exists $C > 0$ such that $C_{\text{obs}}(\gamma) \le C$ for every $\gamma$ in that class.

In the context of feedback stabilization, we mention [2]. For parabolic equations, we also mention [21, 26]. Finally, for the computation of pointwise controls for the Burgers equation, we refer to [4, 31].

The main contributions of this paper are the following. First, we use the HUM method to characterize the control $\widehat{v}$ of minimal $\mathbf{H}'$-norm, also known as the HUM control. We then turn our attention to the numerical approximation of this control and the associated controlled state. Usually (see [16, 29]), such an approximation is computed by minimizing the so-called conjugate functional $\mathcal{J}_\gamma^\star : \mathbf{W} \to \mathbb{R}$ defined by

$$\mathcal{J}_\gamma^\star(\varphi_0,\varphi_1) = \frac{1}{2}\big\|\varphi(\gamma,\cdot)\big\|_{\mathbf{H}}^2 - \int_\Omega y_0\varphi_1 + \langle y_1,\varphi_0\rangle_{-1,1}, \tag{2.4}$$

where $\varphi$ is the solution of (2.3) associated with $(\varphi_0,\varphi_1)$, and $\langle\cdot,\cdot\rangle_{-1,1}$ stands for the duality product in $H_0^1(\Omega)$. Here, instead, we notice that the unconstrained minimization of $\mathcal{J}_\gamma^\star(\varphi_0,\varphi_1)$ is equivalent to the minimization of another functional $\widetilde{\mathcal{J}}_\gamma^\star(\varphi)$ (cf. (2.17)) over $\varphi$ satisfying the constraint $L\varphi = 0$. This constraint is taken into account using a Lagrange multiplier, which leads to a mixed formulation where the space and time variables are embedded. We follow the steps of [9, 13], where a similar formulation is used for controls distributed over noncylindrical domains $q \subset Q_T$. It is worth mentioning that this space-time approach is well adapted to our moving point situation, since we can achieve a good description of the trajectory $\gamma$ embedded in a space-time mesh of $Q_T$. From a numerical point of view, we build a Galerkin approximation of the mixed formulation using conformal space-time finite elements. This allows us to compute the optimal adjoint state $\widehat{\varphi}$, linked to the HUM control $\widehat{v}$ by relation (2.9). This also gives an approximation of the Lagrange multiplier, which turns out to be the controlled state associated with $\widehat{v}$.

Another aspect of this work is the numerical optimization of the support of control. For a given initial datum $(y_0,y_1) \in \mathbf{V}$, we want to minimize the norm $\|\widehat{v}_\gamma\|_{\mathbf{H}'}$ of the HUM control $\widehat{v}_\gamma$ with respect to the trajectory $\gamma$. To do so, we consider the functional

$$J(\gamma) = \frac{1}{2}\|\widehat{v}_\gamma\|_{\mathbf{H}'}^2, \tag{2.5}$$

and we implement a gradient-type algorithm. To find a descent direction at each iteration, we establish a formula for the directional derivative of $J$. The values of $J$ are com-

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

puted using the approximate control arising from the mixed formulation mentioned previously. We perform several numerical experiments and compare our results with those obtained in [6] for controls distributed over noncylindrical domains $q \subset Q_T$. In the simulations the admissible set of trajectories $\gamma$ is discretized using splines functions of degree 5.

The rest of the paper is organized in three sections. First, in Section 2.2, we briefly give some theoretical results. Namely, we justify the existence of weak solutions for the state equation (2.1), and we characterize the control of minimal $\mathbf{H}'$-norm using the HUM method. We also analyze the extremal problem $\min_\gamma J(\gamma)$ (cf. (2.5)) and compute the directional derivative of $J$ with respect to $\gamma$. In a second step, in Section 2.3, we present the space-time mixed formulation used to approximate the control and the controlled state. We also discuss some issues related to the discretization of that formulation. Finally, in Section 2.4, we give several numerical experiments. We illustrate the convergence of the approximated control as the discretization parameter goes to zero. For stationary control points $\gamma \equiv x_0 \in \Omega$, we illustrate the lack of controllability at nonstrategic points. We also describe the gradient-type algorithm designed to optimize the support of control and discuss some results.

## 2.2 Some theoretical results

### 2.2.1 Existence of weak solutions for the state equation

The weak solution of (2.1) is defined by transposition (see [27]). For any $\psi \in L^1(0,T;L^2(\Omega))$, let $\varphi \in C([0,T];H_0^1(\Omega)) \cap C^1([0,T];L^2(\Omega))$ be the solution of the backward adjoint equation

$$L\varphi = \psi \text{ in } Q_T, \quad \varphi = 0 \text{ on } \Sigma_T, \quad (\varphi, \varphi_t)(\cdot, T) = (0,0) \text{ in } \Omega.$$

Multiplying (2.1) by $\varphi$ and integrating by parts, we formally obtain

$$\iint_{Q_T} y\psi = \langle v, \varphi(\gamma,\cdot)\rangle_{\mathbf{H}',\mathbf{H}} - \int_\Omega y_0\varphi_t(\cdot,0) + \langle y_1, \varphi(\cdot,0)\rangle_{-1,1} \quad \forall \psi \in L^1(0,T;L^2(\Omega)), \tag{2.6}$$

where $\langle\cdot,\cdot\rangle_{-1,1}$ and $\langle\cdot,\cdot\rangle_{\mathbf{H}',\mathbf{H}}$ denote, respectively, the duality products in $H_0^1(\Omega)$ and $\mathbf{H}$. We adopt identity (2.6) as the definition of the solution of (2.1) in the sense of transposition. Then we can prove the following result (see [8, Theorem 2.1]).

**Lemma 2.2.1.** *Let $\gamma : [0,T] \to \Omega$ be piecewise $C^1$. If there exists a subdivision $(t_i)_{0\le i\le m}$ of $[0,T]$ such that, on each subinterval $[t_{i-1},t_i]$, $\gamma$ is $C^1$ and $1-|\gamma'|$ does not change sign, then there exists a unique solution $y$ to* (2.1) *in the sense of transposition. This solution has the regularity $y \in C([0,T];L^2(\Omega))$ and $y_t \in L^2([0,T];H^{-1}(\Omega))$.*

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

### 2.2.2 Characterization of the HUM control

To give a characterization of the controls for (2.1), for any $(\varphi_0, \varphi_1) \in \mathbf{W}$, let $\varphi$ be the solution of the adjoint equation (2.3). Multiplying (2.1) by $\varphi$ and integrating by parts, we get that $v \in \mathbf{H}'$ is a control if and only if

$$\langle v, \varphi(\gamma, \cdot)\rangle_{\mathbf{H}',\mathbf{H}} = \int_\Omega y_0\varphi_1 - \langle y_1, \varphi_0\rangle_{-1,1} \quad \forall(\varphi_0, \varphi_1) \in \mathbf{W}. \tag{2.7}$$

Then by a straightforward application of the HUM method (see [8, Section 6]) we can readily characterize the control of minimal $\mathbf{H}'$-norm for (2.1). Let us consider the conjugate functional $\mathcal{J}_\gamma^\star$ defined in (2.4). If $\gamma$ is an admissible trajectory, that is, if the observability inequality (2.2) holds, we can see that $\mathcal{J}_\gamma^\star$ is continuous, strictly convex, and coercive. Thus $\mathcal{J}_\gamma^\star$ has a unique minimum point $(\widehat{\varphi}_0, \widehat{\varphi}_1) \in \mathbf{W}$, which satisfies the optimality condition

$$\langle \widehat{\varphi}(\gamma, \cdot), \varphi(\gamma, \cdot)\rangle_{\mathbf{H}} = \int_\Omega y_0\varphi_1 - \langle y_1, \varphi_0\rangle_{-1,1} \quad \forall(\varphi_0, \varphi_1) \in \mathbf{W}, \tag{2.8}$$

where $\widehat{\varphi}$ and $\varphi$ are the solutions of (2.3) associated with $(\widehat{\varphi}_0, \widehat{\varphi}_1)$ and $(\varphi_0, \varphi_1)$, respectively. For sufficient conditions guaranteeing that a trajectory $\gamma$ is admissible, we refer to [8, Theorem 2.4]. Examples of such admissible trajectories can be found in Figure 2.2 and [8, Section 3]. In view of (2.7), we can then see that the control $\widehat{v}$ of minimal $\mathbf{H}'$-norm for (2.1) has the following form.

**Lemma 2.2.2** (HUM control). *Let $\gamma \in C^1([0,T])$ piecewise. If $\gamma$ is an admissible trajectory, then the control $\widehat{v}$ of minimal $\mathbf{H}'$-norm for* (2.1) *is given by*

$$\begin{aligned}\widehat{v}(t) = &-\frac{\mathrm{d}^2}{\mathrm{d}t^2}\widehat{\varphi}(\gamma(t), t) + \widehat{\varphi}(\gamma(t), t)\\ &+ \frac{\mathrm{d}}{\mathrm{d}t}\widehat{\varphi}(\gamma(t), t)\delta_T(t) - \frac{\mathrm{d}}{\mathrm{d}t}\widehat{\varphi}(\gamma(t), t)\delta_0(t) \quad \forall t \in (0,T),\end{aligned} \tag{2.9}$$

*where $\widehat{\varphi}$ is the solution of* (2.3) *associated with the minimum point $(\widehat{\varphi}_0, \widehat{\varphi}_1)$ of $\mathcal{J}_\gamma^\star$, and $\delta_0$ and $\delta_T$ denote, respectively, the Dirac measures at $t = 0$ and $t = T$. Moreover, the norm of $\widehat{v}$ can be computed by*

$$\|\widehat{v}\|_{\mathbf{H}'}^2 = \|\widehat{\varphi}(\gamma, \cdot)\|_{\mathbf{H}}^2 = \int_0^T \varphi^2(\gamma(t), t)\,\mathrm{d}t + \int_0^T \left|\frac{\mathrm{d}}{\mathrm{d}t}\varphi(\gamma(t), t)\right|^2 \mathrm{d}t. \tag{2.10}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

### 2.2.3 Optimization of the support of control

We focus here on the optimization of the control trajectory. More precisely, for fixed $(y_0, y_1) \in \mathbf{V}$, we want to minimize the norm $\|\widehat{v}\|_{\mathbf{H}'}$ (cf. (2.10)) of the HUM control with respect to the curve $\gamma$, i. e., solve

$$\min_{\gamma \in \mathcal{G}} J(\gamma), \quad \text{where } J(\gamma) = \frac{1}{2}\int_0^T \varphi^2(\gamma(t), t)\,\mathrm{d}t + \frac{1}{2}\int_0^T \left|\frac{\mathrm{d}}{\mathrm{d}t}\varphi(\gamma(t), t)\right|^2 \mathrm{d}t, \tag{2.11}$$

and where $\varphi$ is the solution of (2.3) associated with the minimum point $(\varphi_0, \varphi_1)$ of $\mathcal{J}_\gamma^\star$. The admissible set $\mathcal{G}$ is composed of smooth trajectories, typically, of class $C^2([0, T])$. We also require that the observability inequality (2.2) holds uniformly on $\mathcal{G}$, meaning that there exists $C > 0$ such that $C_{\text{obs}}(\gamma) \leq C$ for every $\gamma \in \mathcal{G}$. This property can be achieved with the hypotheses of [8, Theorem 2.4]. In Section 2.4, we discretize $\mathcal{G}$ using the space $\mathcal{S}_5$ of degree 5 splines, adapted to a fixed regular subdivision of $[0, T]$.

As it stands, we do not know if the extremal problem (2.11) is well posed. To establish the lower semicontinuity of $J$, it could be possible to exploit the works [18, 19], where, in the context of the heat equation, the authors consider a shape optimization problem with respect to a curve. In the process, it might be necessary to have a more regular control, which would probably require more regular initial data $(y_0, y_1)$ (see [15]).

Moreover, a longer trajectory $\gamma$ allows intuitively a smaller cost of control. Consequently, to give more sense to the problem, we penalize the length $L(\gamma)$ of the curve $\gamma$. Similarly, to avoid too fast variations of the trajectory, we also regularize the "curvature" $\gamma''$. A similar strategy has been introduced and discussed in [6]. Thus, for $\varepsilon > 0$ small enough, $\eta > 0$ large enough, and $\overline{L} \geq T$ fixed, we consider the following regularized-penalized extremal problem:

$$\min_{\gamma \in \mathcal{G}} J_{\varepsilon,\eta}(\gamma), \quad \text{where } J_{\varepsilon,\eta}(\gamma) = J(\gamma) + \frac{\varepsilon}{2}\|\gamma''\|^2_{L^2(0,T)} + \frac{\eta}{2}\left((L(\gamma) - \overline{L})^+\right)^2, \tag{2.12}$$

and where $(\cdot)^+$ stands for the positive part.

We solve this problem numerically in Section 2.4 using a gradient-type algorithm. To evaluate a descent direction for $J_{\varepsilon,\eta}$ at each iteration of the algorithm, we compute the derivatives of $J$ and $J_{\varepsilon,\eta}$ with respect to $\gamma$.

**Lemma 2.2.3.** *Let $\gamma \in C^2([0, T])$ be an admissible trajectory, and let $\overline{\gamma} \in C^2([0, T])$ be a perturbation. The directional derivative of $J$ at $\gamma$ in the direction $\overline{\gamma}$, defined by* $\mathrm{d}J(\gamma; \overline{\gamma}) :=$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$\lim_{\nu\to 0} \frac{J(\gamma+\nu\overline{\gamma})-J(\gamma)}{\nu}$, *reads as follows:*

$$\begin{aligned} \mathrm{d}J(\gamma;\overline{\gamma}) = &-\int_0^T \varphi(\gamma(t),t)\varphi_x(\gamma(t),t)\overline{\gamma}(t)\,\mathrm{d}t \\ &-\int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\varphi(\gamma(t),t)\frac{\mathrm{d}}{\mathrm{d}t}(\varphi_x(\gamma(t),t)\overline{\gamma}(t))\mathrm{d}t, \end{aligned}$$

*where* $\varphi$ *is the solution of* (2.3) *associated with the minimum point* $(\varphi_0,\varphi_1)$ *of* $\mathcal{J}_\gamma^\star$. *Similarly, the directional derivative of* $J_{\varepsilon,\eta}$ *at* $\gamma$ *in the direction* $\overline{\gamma}$ *is given by*

$$\mathrm{d}J_{\varepsilon,\eta}(\gamma;\overline{\gamma}) = \mathrm{d}J(\gamma;\overline{\gamma}) + \varepsilon\langle\gamma'',\overline{\gamma}''\rangle_{L^2(0,T)} + \eta\big(L(\gamma)-\overline{L}\big)^+\mathrm{d}L(\gamma;\overline{\gamma}),$$

*where*

$$L(\gamma) = \int_0^T \sqrt{1+\gamma'^2} \quad \textit{and} \quad \mathrm{d}L(\gamma;\overline{\gamma}) = \int_0^T \frac{\gamma'}{\sqrt{1+\gamma'^2}}\overline{\gamma}'.$$

*Proof.* We provide only a formal proof. Rigorous demonstrations of similar lemmas can be found in [6, 30] for controls distributed over domains $q \subset Q_T$. For any admissible trajectory $\gamma \in C^2([0,T])$ and any perturbation $\overline{\gamma} \in C^2([0,T])$, we get

$$\begin{aligned} \mathrm{d}J(\gamma;\overline{\gamma}) = &\int_0^T \varphi(\gamma(t),t)(\varphi'(\gamma(t),t)+\varphi_x(\gamma(t),t)\overline{\gamma}(t))\mathrm{d}t \\ &+\int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\varphi(\gamma(t),t)\frac{\mathrm{d}}{\mathrm{d}t}(\varphi'(\gamma(t),t)+\varphi_x(\gamma(t),t)\overline{\gamma}(t))\mathrm{d}t. \end{aligned} \tag{2.13}$$

Here $\varphi'$ denotes the derivative of $\varphi$ with respect to $\gamma$. To simplify (2.13), we differentiate the optimality condition (2.8) with respect to $\gamma$. This gives

$$\begin{aligned} &\int_0^T (\varphi'(\gamma(t),t)+\varphi_x(\gamma(t),t)\overline{\gamma}(t))\psi(\gamma(t),t)\,\mathrm{d}t + \int_0^T \varphi(\gamma(t),t)\psi_x(\gamma(t),t)\overline{\gamma}(t)\,\mathrm{d}t \\ &\quad+\int_0^T \frac{\mathrm{d}}{\mathrm{d}t}(\varphi'(\gamma(t),t)+\varphi_x(\gamma(t),t)\overline{\gamma}(t))\frac{\mathrm{d}}{\mathrm{d}t}\psi(\gamma(t),t)\,\mathrm{d}t \\ &\quad+\int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\varphi(\gamma(t),t)\frac{\mathrm{d}}{\mathrm{d}t}(\psi_x(\gamma(t),t)\overline{\gamma}(t))\mathrm{d}t = 0 \quad \forall(\psi_0,\psi_1)\in\mathbf{W}, \end{aligned}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

where $\psi$ is the solution of (2.3) associated with $(\psi_0, \psi_1)$. Evaluating the previous expression for $(\psi_0, \psi_1) = (\varphi_0, \varphi_1)$, we can eliminate the derivative $\varphi'$ from (2.13) and obtain the announced result. □

## 2.3 Mixed formulation

In this section, to approximate the HUM control for (2.1) and the associated controlled state, we present a space-time mixed formulation based on the optimality condition (2.8). We follow the steps of [9, Section 3.1], where a similar formulation is built for controls distributed over domains $q \subset Q_T$. From a numerical point of view, this space-time formulation is very appropriate for the moving point situation considered in this work. Indeed, after the discretization step, we solve the formulation using a space-time triangular mesh, which is constructed from boundary vertices placed on the border of $Q_T$ and on the curve $\gamma$.

### 2.3.1 Mixed formulation

We start by a lemma extending the observability inequality (2.2). For this, we first need to introduce the functional space

$$\mathbf{\Phi} := \{\varphi \in C([0,T]; H_0^1(\Omega)) \cap C^1([0,T]; L^2(\Omega));\ L\varphi \in L^2(0,T; L^2(\Omega))\}.$$

**Lemma 2.3.1** (Generalized observability inequality). *Let $\gamma \in C^1([0,T])$ piecewise. If $\gamma$ is an admissible trajectory, there exists a constant $\widetilde{C}_{\text{obs}}(\gamma) > 0$ such that*

$$\|(\varphi, \varphi_t)(\cdot, 0)\|_{\mathbf{W}}^2 \leq \widetilde{C}_{\text{obs}}(\gamma)\left(\|\varphi(\gamma, \cdot)\|_{\mathbf{H}}^2 + \|L\varphi\|_{L^2(0,T;L^2(\Omega))}^2\right) \quad \forall \varphi \in \mathbf{\Phi}. \tag{2.14}$$

*Proof.* Let $\varphi \in \mathbf{\Phi}$. We can decompose $\varphi = \psi_1 + \psi_2$, where $\psi_1, \psi_2 \in \mathbf{\Phi}$ solve

$$\begin{cases} L\psi_1 = 0 \text{ in } Q_T, \quad \psi_1 = 0 \text{ on } \Sigma_T, \quad (\psi_1, \psi_{1,t})(\cdot, 0) = (\varphi, \varphi_t)(\cdot, 0) \text{ in } \Omega, \\ L\psi_2 = L\varphi \text{ in } Q_T, \quad \psi_2 = 0 \text{ on } \Sigma_T, \quad (\psi_2, \psi_{2,t})(\cdot, 0) = (0, 0) \text{ in } \Omega. \end{cases}$$

From Duhamel's principle and the conservation of energy we can show (see [8, Section 5]) the following so-called hidden regularity property for $\psi_2$: there exists a constant $c(\gamma) > 0$ such that

$$\|\psi_2(\gamma, \cdot)\|_{\mathbf{H}}^2 \leq c(\gamma)\|L\varphi\|_{L^2(0,T;L^2(\Omega))}^2. \tag{2.15}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Combining (2.2) for $\psi_1$ and (2.15) for $\psi_2$, we obtain

$$\begin{aligned}\|(\varphi,\varphi_t)(\cdot,0)\|_{\mathbf{W}}^2 = \|(\psi_1,\psi_{1,t})(\cdot,0)\|_{\mathbf{W}}^2 &\le C_{\mathrm{obs}}(\gamma)\|\psi_1(\gamma,\cdot)\|_{\mathbf{H}}^2\\ &\le 2C_{\mathrm{obs}}(\gamma)(\|\varphi(\gamma,\cdot)\|_{\mathbf{H}}^2 + \|\psi_2(\gamma,\cdot)\|_{\mathbf{H}}^2)\\ &\le \widetilde{C}_{\mathrm{obs}}(\gamma)(\|\varphi(\gamma,\cdot)\|_{\mathbf{H}}^2 + \|L\varphi\|_{L^2(0,T;L^2(\Omega))}^2).\end{aligned}$$

□

As for (2.2), it is possible to find a class of admissible trajectories $\gamma$ such that the generalized observability inequality (2.14) holds uniformly (see [8, Theorem 2.4]), i. e., there exists $\widetilde{C} > 0$ such that $\widetilde{C}_{\mathrm{obs}}(\gamma) \le \widetilde{C}$ for every $\gamma$ in that class. In addition, inequality (2.14) implies the following property on the space $\mathbf{\Phi}$.

**Lemma 2.3.2.** *Let $\gamma \in C^1([0,T])$ piecewise. If $\gamma$ is an admissible trajectory, then $\mathbf{\Phi}$ is a Hilbert space with the inner product*

$$\langle\varphi,\overline{\varphi}\rangle_{\mathbf{\Phi}} = \big\langle\varphi(\gamma,\cdot),\overline{\varphi}(\gamma,\cdot)\big\rangle_{\mathbf{H}} + \tau\langle L\varphi, L\overline{\varphi}\rangle_{L^2(0,T;L^2(\Omega))} \quad \forall\varphi,\overline{\varphi}\in\mathbf{\Phi} \tag{2.16}$$

*with fixed $\tau > 0$.*

*Proof.* The sem-norm $\|\cdot\|_{\mathbf{\Phi}}$ associated with the inner product is trivially a norm in view of the generalized observability inequality (2.14). It remains to prove that $\mathbf{\Phi}$ is complete with respect to this norm. Let $(\varphi_k)_{k\ge1} \subset \mathbf{\Phi}$ be a Cauchy sequence for the norm $\|\cdot\|_{\mathbf{\Phi}}$. So there exists $f \in L^2(0,T;L^2(\Omega))$ such that $L\varphi_k \to f$ in $L^2(0,T;L^2(\Omega))$. As a consequence of (2.14), there also exists $(\varphi_0,\varphi_1) \in \mathbf{W}$ such that $(\varphi_k,\varphi_{k,t})(\cdot,0) \to (\varphi_0,\varphi_1)$ in $\mathbf{W}$. Therefore $(\varphi_k)_{k\ge1}$ can be considered as a sequence of solutions of the wave equation with convergent initial data and convergent right-hand sides. By the continuous dependence of the solution of the wave equation on the data, $\varphi_k \to \varphi$ in $C([0,T];H_0^1(\Omega)) \cap C^1([0,T];L^2(\Omega))$, where $\varphi$ is the solution of the wave equation with initial datum $(\varphi_0,\varphi_1) \in \mathbf{W}$ and right-hand side $f \in L^2(0,T;L^2(\Omega))$. Thus $\varphi \in \mathbf{\Phi}$. □

We can now turn to the setup of the mixed formulation. To avoid the minimization of the conjugate functional $\mathcal{J}_\gamma^\star$ (cf. (2.4)) with respect to $(\varphi_0,\varphi_1)$, we remark that the solution $\varphi$ of (2.3) is completely and uniquely determined by the initial datum $(\varphi_0,\varphi_1)$. Then the main idea of the reformulation is to keep $\varphi$ as a main variable and consider instead the minimization of

$$\widetilde{\mathcal{J}}_\gamma^\star(\varphi) = \frac{1}{2}\|\varphi(\gamma,\cdot)\|_{\mathbf{H}}^2 - \int_\Omega y_0\varphi_t(\cdot,0) + \big\langle y_1,\varphi(\cdot,0)\big\rangle_{-1,1} \tag{2.17}$$

over

$$\mathbf{\Phi}_0 := \{\varphi\in\mathbf{\Phi};\ L\varphi = 0 \in L^2(0,T;L^2(\Omega))\}.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Indeed, we clearly have

$$\min_{(\varphi_0,\varphi_1)\in\mathbf{W}} \mathcal{J}_\gamma^\star(\varphi_0,\varphi_1) = \mathcal{J}_\gamma^\star(\widehat{\varphi}_0,\widehat{\varphi}_1) = \widetilde{\mathcal{J}}_\gamma^\star(\widehat{\varphi}) = \min_{\varphi\in\mathbf{\Phi}_0} \widetilde{\mathcal{J}}_\gamma^\star(\varphi),$$

where $\widehat{\varphi}$ is the solution of (2.3) associated with the minimum point $(\widehat{\varphi}_0,\widehat{\varphi}_1)$ of $\mathcal{J}_\gamma^\star$. Besides, the minimum point $\widehat{\varphi}$ of $\widetilde{\mathcal{J}}_\gamma^\star$ is unique. So the new variable is the function $\varphi$ with the constraint $L\varphi = 0$ in $L^2(0,T;L^2(\Omega))$. To deal with this constraint, we introduce a Lagrange multiplier $\lambda \in \mathbf{\Lambda} := L^2(0,T;L^2(\Omega))$. We thus consider the following problem: find a solution $(\varphi,\lambda) \in \mathbf{\Phi}\times\mathbf{\Lambda}$ of

$$\begin{cases} a(\varphi,\overline{\varphi}) - b(\overline{\varphi},\lambda) = \ell(\overline{\varphi}) & \forall \overline{\varphi} \in \mathbf{\Phi}, \\ b(\varphi,\overline{\lambda}) = 0 & \forall \overline{\lambda} \in \mathbf{\Lambda}, \end{cases} \tag{2.18}$$

where we have set

$$\begin{aligned} &a : \mathbf{\Phi}\times\mathbf{\Phi} \to \mathbb{R}, \quad a(\varphi,\overline{\varphi}) = \big\langle \varphi(\gamma,\cdot), \overline{\varphi}(\gamma,\cdot)\big\rangle_{\mathbf{H}}, \\ &b : \mathbf{\Phi}\times\mathbf{\Lambda} \to \mathbb{R}, \quad b(\varphi,\lambda) = \langle L\varphi,\lambda\rangle_{L^2(0,T;L^2(\Omega))}, \\ &\ell : \mathbf{\Phi} \to \mathbb{R}, \quad \ell(\varphi) = \int_\Omega y_0\varphi_t(\cdot,0) - \big\langle y_1, \varphi(\cdot,0)\big\rangle_{-1,1}. \end{aligned}$$

The introduction of this problem is justified by the following result.

**Theorem 2.3.1** (Mixed formulation). *Let $\gamma \in C^1([0,T])$ piecewise. If $\gamma$ is an admissible trajectory, then we have the following properties;*

- *The mixed formulation* (2.18) *is well posed.*
- *The unique solution $(\varphi,\lambda) \in \mathbf{\Phi}\times\mathbf{\Lambda}$ is the unique saddle point of the Lagrangian $\mathcal{L} : \mathbf{\Phi}\times\mathbf{\Lambda} \to \mathbb{R}$ defined by*

$$\mathcal{L}(\varphi,\lambda) = \frac{1}{2}a(\varphi,\varphi) - b(\varphi,\lambda) - \ell(\varphi).$$

- *The optimal function $\varphi$ is the minimum point of $\widetilde{\mathcal{J}}_\gamma^\star$ over $\mathbf{\Phi}_0$. Besides, the optimal function $\lambda \in \mathbf{\Lambda}$ is the solution of the controlled wave equation* (2.1) *with the control $v$ associated with $\varphi$ (cf.* (2.9)*).*

*Proof.* We easily check that the bilinear form $a$ is continuous over $\mathbf{\Phi}\times\mathbf{\Phi}$, symmetric and positive. Similarly, we check that the bilinear form $b$ is continuous over $\mathbf{\Phi}\times\mathbf{\Lambda}$. Furthermore, the continuity of the linear form $\ell$ over $\mathbf{\Phi}$ is a direct consequence of the generalized observability inequality (2.14),

$$\big|\ell(\varphi)\big| \le \big\|(y_0,y_1)\big\|_{\mathbf{V}} \sqrt{2\widetilde{C}_{\mathrm{obs}}(\gamma)\max(1,\tau^{-1})}\,\|\varphi\|_{\mathbf{\Phi}} \quad \forall \varphi \in \mathbf{\Phi}.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Therefore, to prove the well-posedness of the mixed formulation (2.18), we only need to check the following two properties (see [7]).

- The form $a$ is coercive on the kernel $\mathcal{N}(b) := \{\varphi \in \mathbf{\Phi};\ b(\varphi,\lambda) = 0,\ \forall \lambda \in \mathbf{\Lambda}\}$.
- The form $b$ satisfies the usual "inf–sup" condition over $\mathbf{\Phi} \times \mathbf{\Lambda}$, i. e., there exists a constant $\delta > 0$ such that

$$\inf_{\lambda\in\mathbf{\Lambda}} \sup_{\varphi\in\mathbf{\Phi}} \frac{b(\varphi,\lambda)}{\|\varphi\|_{\mathbf{\Phi}}\|\lambda\|_{\mathbf{\Lambda}}} \geq \delta. \tag{2.19}$$

The first point is clear from the definition of $a$. Indeed, for any $\varphi \in \mathcal{N}(b) = \mathbf{\Phi}_0$, $a(\varphi,\varphi) = \|\varphi\|^2_{\mathbf{\Phi}}$. We now check the inf–sup condition (2.19). For any $\lambda_0 \in \mathbf{\Lambda}$, we define the unique element $\varphi_0 \in \mathbf{\Phi}$ such that

$$L\varphi_0 = \lambda_0 \text{ in } Q_T, \quad \varphi_0 = 0 \text{ on } \Sigma_T, \quad (\varphi_0, \varphi_{0,t})(\cdot,0) = (0,0) \text{ in } \Omega.$$

This implies $b(\varphi_0,\lambda_0) = \|\lambda_0\|^2_{\mathbf{\Lambda}}$ and

$$\sup_{\varphi\in\mathbf{\Phi}} \frac{b(\varphi,\lambda_0)}{\|\varphi\|_{\mathbf{\Phi}}\|\lambda_0\|_{\mathbf{\Lambda}}} \geq \frac{b(\varphi_0,\lambda_0)}{\|\varphi_0\|_{\mathbf{\Phi}}\|\lambda_0\|_{\mathbf{\Lambda}}} = \frac{\|\lambda_0\|_{\mathbf{\Lambda}}}{\sqrt{\|\varphi_0(\gamma,\cdot)\|^2_{\mathbf{H}} + \tau\|\lambda_0\|^2_{\mathbf{\Lambda}}}}.$$

We then use the following estimate (see [8, Section 5]): there exists a constant $c(\gamma) > 0$ such that

$$\|\varphi_0(\gamma,\cdot)\|^2_{\mathbf{H}} \leq c(\gamma)\|\lambda_0\|^2_{\mathbf{\Lambda}}.$$

Combining the two previous inequalities, we obtain

$$\sup_{\varphi\in\mathbf{\Phi}} \frac{b(\varphi,\lambda_0)}{\|\varphi\|_{\mathbf{\Phi}}\|\lambda_0\|_{\mathbf{\Lambda}}} \geq \frac{1}{\sqrt{c(\gamma)+\tau}} \quad \forall \lambda_0 \in \mathbf{\Lambda}.$$

Hence inequality (2.19) holds with $\delta = (c(\gamma)+\tau)^{-\frac{1}{2}}$.

The second point of the theorem is due to the symmetry and positivity of the bilinear form $a$. Regarding the third point, the equality $b(\varphi,\overline{\lambda}) = 0$ for all $\overline{\lambda} \in \mathbf{\Lambda}$ implies that $L\varphi = 0$ in $L^2(0,T;L^2(\Omega))$. Besides, for $\overline{\varphi} \in \mathbf{\Phi}_0$, the first equation of (2.18) gives $a(\varphi,\overline{\varphi}) = \ell(\overline{\varphi})$. So, if $(\varphi,\lambda) \in \mathbf{\Phi} \times \mathbf{\Lambda}$ solves the mixed formulation, then $\varphi \in \mathbf{\Phi}_0$ and $\mathcal{L}(\varphi,\lambda) = \widetilde{\mathcal{J}}^\star_\gamma(\varphi)$. Moreover, again due to the symmetry and positivity of $a$, the function $\varphi$ is the minimum point of $\widetilde{\mathcal{J}}^\star_\gamma$ over $\mathbf{\Phi}_0$. Indeed, for any $\overline{\varphi} \in \mathbf{\Phi}_0$, we have

$$\widetilde{\mathcal{J}}^\star_\gamma(\varphi) = -\frac{1}{2}a(\varphi,\varphi) \leq \frac{1}{2}a(\overline{\varphi},\overline{\varphi}) - a(\varphi,\overline{\varphi}) = \frac{1}{2}a(\overline{\varphi},\overline{\varphi}) - \ell(\overline{\varphi}) = \widetilde{\mathcal{J}}^\star_\gamma(\overline{\varphi}).$$

Finally, the first equation of (2.18) reads

$$\langle \varphi(\gamma,\cdot), \overline{\varphi}(\gamma,\cdot)\rangle_{\mathbf{H}} - \langle L\overline{\varphi},\lambda\rangle_{\mathbf{\Lambda}} = \int_\Omega y_0\overline{\varphi}_t(\cdot,0) - \langle y_1, \overline{\varphi}(\cdot,0)\rangle_{-1,1} \quad \forall \overline{\varphi} \in \mathbf{\Phi}.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Since the control $v$ of minimal $\mathbf{H}'$-norm is given by (2.9), we get

$$\iint_{Q_T} \lambda L\overline{\varphi} = \langle v, \overline{\varphi}(\gamma, \cdot)\rangle_{\mathbf{H}',\mathbf{H}} - \int_\Omega y_0 \overline{\varphi}_t(\cdot, 0) + \langle y_1, \overline{\varphi}(\cdot, 0)\rangle_{-1,1} \quad \forall \overline{\varphi} \in \mathbf{\Phi},$$

which means that $\lambda$ is solution in a weak sense of the wave equation (2.1) associated with the initial datum $(y_0, y_1) \in \mathbf{V}$ and the control $v \in \mathbf{H}'$. □

Consequently, the search of the HUM control for (2.1) is reduced to the resolution of the mixed formulation (2.18) or, equivalently, to the search of the saddle point of $\mathcal{L}$. Moreover, for numerical purposes, it is convenient to "augment" the Lagrangian $\mathcal{L}$ and to consider instead the Lagrangian $\mathcal{L}_r$ defined, for any $r > 0$, by

$$\begin{cases} \mathcal{L}_r(\varphi, \lambda) = \frac{1}{2} a_r(\varphi, \varphi) - b(\varphi, \lambda) - \ell(\varphi), \\ a_r(\varphi, \overline{\varphi}) = a(\varphi, \overline{\varphi}) + r\langle L\varphi, L\overline{\varphi}\rangle_{L^2(0,T;L^2(\Omega))}. \end{cases}$$

Since $a(\varphi, \varphi) = a_r(\varphi, \varphi)$ for $\varphi \in \mathbf{\Phi}_0$, the Lagrangians $\mathcal{L}$ and $\mathcal{L}_r$ share the same saddle point.

## 2.3.2 Discretization

We now turn to the discretization of the mixed formulation (2.18). Let $(\mathbf{\Phi}_h)_{h>0} \subset \mathbf{\Phi}$ and $(\mathbf{\Lambda}_h)_{h>0} \subset \mathbf{\Lambda}$ be two families of finite-dimensional spaces. For any $h > 0$, we introduce the following approximated problem: find a solution $(\varphi_h, \lambda_h) \in \mathbf{\Phi}_h \times \mathbf{\Lambda}_h$ of

$$\begin{cases} a_r(\varphi_h, \overline{\varphi}_h) - b(\overline{\varphi}_h, \lambda_h) = \ell(\overline{\varphi}_h) & \forall \overline{\varphi}_h \in \mathbf{\Phi}_h, \\ b(\varphi_h, \overline{\lambda}_h) = 0 & \forall \overline{\lambda}_h \in \mathbf{\Lambda}_h. \end{cases} \tag{2.20}$$

To prove the well-posedness of this mixed formulation, we again have to check the following two properties. First, the bilinear form $a_r$ is coercive on the kernel $\mathcal{N}_h(b) := \{\varphi_h \in \mathbf{\Phi}_h;\ b(\varphi_h, \lambda_h) = 0,\ \forall \lambda_h \in \mathbf{\Lambda}_h\}$. Indeed, from the relation

$$a_r(\varphi, \varphi) \geq \min(1, r/\tau)\|\varphi\|^2_{\mathbf{\Phi}} \quad \forall \varphi \in \mathbf{\Phi}$$

it follows that the form $a_r$ is coercive on the full space $\mathbf{\Phi}$ and so a fortiori on $\mathcal{N}_h(b) \subset \mathbf{\Phi}_h \subset \mathbf{\Phi}$. The second property is a discrete inf–sup condition: there exists a constant $\delta_h > 0$ such that

$$\inf_{\lambda_h \in \mathbf{\Lambda}_h} \sup_{\varphi_h \in \mathbf{\Phi}_h} \frac{b(\varphi_h, \lambda_h)}{\|\varphi_h\|_{\mathbf{\Phi}_h} \|\lambda_h\|_{\mathbf{\Lambda}_h}} \geq \delta_h. \tag{2.21}$$

The spaces $\mathbf{\Phi}_h$ and $\mathbf{\Lambda}_h$ are finite-dimensional, so the infimum and supremum in (2.21) are reached. Moreover, from the properties of $a_r$ and with the finite element spaces

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$\mathbf{\Phi}_h$, $\mathbf{\Lambda}_h$ chosen below, it is standard to prove that $\delta_h$ is strictly positive. Consequently, for any $h > 0$, there exists a unique couple $(\varphi_h, \lambda_h) \in \mathbf{\Phi}_h \times \mathbf{\Lambda}_h$ solution of the discrete mixed formulation (2.21).

On the other hand, if we could show that $\inf_{h>0} \delta_h > 0$, then it would ensure the convergence of the solution $(\varphi_h, \lambda_h)$ of the discrete formulation (2.20) toward the solution $(\varphi, \lambda)$ of the continuous formulation (2.18). However, this property is usually difficult to prove and depends strongly on the choice made for the spaces $\mathbf{\Phi}_h$, $\mathbf{\Lambda}_h$. We analyze numerically this property in Section 2.3.3.

Let us consider a triangulation $\mathcal{T}_h$ of $Q_T$, i. e., $\bigcup_{K\in\mathcal{T}_h} K = \overline{Q_T}$. We denote $h := \max\{\mathrm{diam}(K);\ K \in \mathcal{T}_h\}$, where $\mathrm{diam}(K)$ is the diameter of the triangle $K$. In what follows, the space-time mesh $\mathcal{T}_h$ is built from a discretization of the border of $Q_T$ and the curve $\gamma$ (see Figure 2.5). Thus the fineness of $\mathcal{T}_h$ will be given either by $h$ or by the number $N_{\mathcal{T}}$ of vertices per unit of length. This also means that some vertices are supported on $\gamma$, making the mesh well adapted to the control trajectory. The mesh is generated using the software FreeFEM++ (see [17]).

The finite-dimensional space $\mathbf{\Phi}_h$ must be chosen such that $L\varphi_h$ belongs to $L^2(0,T;L^2(\Omega))$ for all $\varphi_h \in \mathbf{\Phi}_h$. Therefore any space of functions continuously differentiable with respect to both $x$ and $t$ is a conformal approximation of $\mathbf{\Phi}$. We define the space $\mathbf{\Phi}_h$ as follows:

$$\mathbf{\Phi}_h := \{\varphi_h \in C^1(\overline{Q_T});\ \varphi_{h|_K} \in \mathbb{P}(K),\ \forall K \in \mathcal{T}_h,\ \varphi_h = 0 \text{ on } \Sigma_T\} \subset \mathbf{\Phi},$$

where $\mathbb{P}(K)$ stands for the complete *Hsieh–Clough–Tocher* (HCT) finite element of class $C^1$. It is a so-called composite finite element. It involves 12 degrees of freedom, which are, for each triangle $K$, the values of $\varphi_h$, $\varphi_{h,x}$, $\varphi_{h,t}$ on the three vertices and the values of the normal derivative of $\varphi$ in the middle of the three edges. We refer to [12] and [5, 28] for the precise definition and the implementation of such a finite element. We also introduce the finite-dimensional space

$$\mathbf{\Lambda}_h := \{\lambda_h \in C^0(\overline{Q_T});\ \lambda_{h|_K} \in \mathbb{Q}(K),\ \forall K \in \mathcal{T}_h,\ \lambda_h = 0 \text{ on } \Sigma_T\} \subset \mathbf{\Lambda},$$

where $\mathbb{Q}(K)$ is the space of affine functions in both $x$ and $t$ on the element $K$.

Let $n_h = \dim(\mathbf{\Phi}_h)$ and $m_h = \dim(\mathbf{\Lambda}_h)$. We define the matrices $A_{r,h} \in \mathbb{R}^{n_h,n_h}$, $B_h \in \mathbb{R}^{m_h,n_h}$, and $M_h \in \mathbb{R}^{m_h,m_h}$ and the vector $L_h \in \mathbb{R}^{n_h}$ by

$$\begin{aligned}
\langle A_{r,h}\{\varphi_h\}, \{\overline{\varphi}_h\}\rangle &= a_r(\varphi_h, \overline{\varphi}_h) \quad \forall \varphi_h, \overline{\varphi}_h \in \mathbf{\Phi}_h,\\
\langle B_h\{\varphi_h\}, \{\lambda_h\}\rangle &= b(\varphi_h, \lambda_h) \quad \forall \varphi_h \in \mathbf{\Phi}_h, \forall \lambda_h \in \mathbf{\Lambda}_h,\\
\langle M_h\{\lambda_h\}, \{\overline{\lambda}_h\}\rangle &= \langle \lambda_h, \overline{\lambda}_h\rangle_{\mathbf{\Lambda}} \quad \forall \lambda_h, \overline{\lambda}_h \in \mathbf{\Lambda}_h,\\
\langle L_h, \{\varphi_h\}\rangle &= \ell(\varphi_h) \quad \forall \varphi_h \in \mathbf{\Phi}_h,
\end{aligned}$$

where $\{\varphi_h\} \in \mathbb{R}^{n_h}$ and $\{\lambda_h\} \in \mathbb{R}^{m_h}$ are the vectors associated with $\varphi_h \in \mathbf{\Phi}_h$ and $\lambda_h \in \mathbf{\Lambda}_h$, respectively. With these notations, the discrete mixed formulation (2.20) reads as

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

follows: find $\{\varphi_h\} \in \mathbb{R}^{n_h}$ and $\{\lambda_h\} \in \mathbb{R}^{m_h}$ such that

$$\begin{pmatrix} A_{r,h} & -B_h^T \\ -B_h & 0 \end{pmatrix} \begin{pmatrix} \{\varphi_h\} \\ \{\lambda_h\} \end{pmatrix} = \begin{pmatrix} L_h \\ 0 \end{pmatrix}. \tag{2.22}$$

For any $r > 0$, the matrix $A_{r,h}$ is symmetric and positive definite. However, the matrix in (2.22) is symmetric but not positive definite. System (2.22) is solved by the LU method with FreeFEM++ (see [17]).

### 2.3.3 Discrete inf–sup test

Here we numerically test the discrete inf–sup condition (2.21) and more precisely the property $\inf_{h>0} \delta_h > 0$. For simplicity, we take $\tau = r > 0$ in (2.16), so that $a_{r,h}(\varphi, \overline{\varphi}) = \langle \varphi, \overline{\varphi} \rangle_{\mathbf{\Phi}}$ for all $\varphi, \overline{\varphi} \in \mathbf{\Phi}$. It is readily seen (see [10]) that the discrete inf–sup constant satisfies

$$\delta_h = \inf\{\sqrt{\mu};\ B_h A_{r,h}^{-1} B_h^T \{\lambda_h\} = \mu M_h \{\lambda_h\},\ \forall \{\lambda_h\} \in \mathbb{R}^{m_h} \setminus \{0\}\}. \tag{2.23}$$

For any $h > 0$, the matrix $B_h A_{r,h}^{-1} B_h^T$ is symmetric and positive definite, so the constant $\delta_h$ is strictly positive. The generalized eigenvalue problem (2.23) is solved by the inverse power method (see [11]). Given $\{u_h^0\} \in \mathbb{R}^{m_h}$ such that $\|\{u_h^0\}\|_2 = 1$ for any $n \in \mathbb{N}$, compute iteratively $(\{\varphi_h^n\}, \{\lambda_h^n\}) \in \mathbb{R}^{n_h} \times \mathbb{R}^{m_h}$ and $\{u_h^{n+1}\} \in \mathbb{R}^{m_h}$ as follows:

$$\begin{cases} A_{r,h}\{\varphi_h^n\} - B_h^T\{\lambda_h^n\} = 0, \\ B_h\{\varphi_h^n\} = M_h\{u_h^n\}, \end{cases} \qquad \{u_h^{n+1}\} = \frac{\{\lambda_h^n\}}{\|\{\lambda_h^n\}\|_2}.$$

The discrete inf–sup constant $\delta_h$ is then given by $\delta_h = \lim_{n\to\infty} \|\{\lambda_h^n\}\|_2^{-\frac{1}{2}}$.

We now compute $\delta_h$ for decreasing values of the fineness $h$ and for different values of the parameter $r$, namely, $r = 10^{-2}$, $r = h$, and $r = h^2$. We use the control trajectory $y$ defined in (**Ex1**–$y$). The values we obtain are collected in Table 2.1. In view of the results for $r = 10^{-2}$, the constant $\delta_h$ does not seem to be uniformly bounded below as $h \to 0$. Thus we may conclude that the finite elements used here do not "pass" the discrete inf–sup test. As we will see in the next section, this fact does not prevent the convergence of the sequences $(\varphi_h)_{h>0}$ and $(\lambda_h)_{h>0}$, at least for the cases we have considered. We also observe that $\delta_h$ remains bounded below with respect to $h$ when $r$ depends appropriately on $h$, as, for instance, in the case $r = h^2$.

**Table 2.1:** Discrete inf–sup constant $\delta_h$ w. r. t. $h$ and $r$, for $y$ defined in (**Ex1**–$y$).

| $h$ ($\times 10^{-2}$) | 6.46 | 3.51 | 2.66 | 2.17 | 1.37 | 1.21 |
|---|---|---|---|---|---|---|
| $r = 10^{-2}$ | 1.8230 | 1.7947 | 1.7845 | 1.6749 | 1.6060 | 1.5008 |
| $r = h$ | 1.4575 | 1.3806 | 1.3269 | 1.2402 | 1.4188 | 1.3851 |
| $r = h^2$ | 1.8873 | 1.8885 | 1.8783 | 1.8697 | 1.8982 | 1.8920 |

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

## 2.4 Numerical simulations

In this section, we solve on various examples the discrete mixed formulation (2.20) to compute the HUM control for (2.1) and the associated controlled state. First, we determine the rate of convergence of the approximated control/controlled state as the discretization parameter $h$ goes to zero. Second, for stationary control points $\gamma \equiv x_0$, we illustrate the blowup of the cost of control at nonstrategic points. Finally, we introduce a gradient-type algorithm to solve problem (2.12) of optimizing the support of control. The algorithm is then tested on two different initial data. From now on, we set $T = 2$ and $r = 10^{-2}$.

### 2.4.1 Convergence of the approximated control

To measure the rate of convergence of the approximated control with respect to the mesh fineness $h$, we use the initial datum

$$y_0(x) = \sin(\pi x), \quad y_1(x) = 0, \quad x \in \Omega, \tag{\textbf{Ex1}–$\mathbf{y}_0$}$$

and the control trajectory

$$\gamma(t) = \frac{1}{5} + \frac{3}{5}\frac{t}{T}, \quad t \in [0, T]. \tag{\textbf{Ex1}–$\gamma$}$$

This curve $\gamma$ is an admissible trajectory (see [8, Example 3.2]), i. e., system (2.1) is controllable. To compare with the approximated solution $(\varphi_h, \lambda_h)$ of (2.20), using the optimality condition (2.8), we compute another approximation $(\varphi, \lambda)$ by Fourier expansion (see the Appendix) with $N_F = 100$ harmonics. We then evaluate the errors $\|\varphi(\gamma, \cdot) - \varphi_h(\gamma, \cdot)\|_{\mathbf{H}}$ and $\|\lambda - \lambda_h\|_{\mathbf{\Lambda}}$ for the six fineness levels $N_{\mathcal{T}} = 25, 50, 75, 100, 125, 150$. We gather the results in Table 2.2 and display them in Figure 2.1. By linear regression we find a convergence rate in $h^{0.44}$ for $\varphi_h$ and in $h^{0.48}$ for $\lambda_h$. In Figure 2.2, we represent the adjoint state $\varphi_h$ and the controlled state $\lambda_h$ for $N_{\mathcal{T}} = 150$. The HUM control $v_h$ computed from $\varphi_h$ by (2.9) is shown in Figure 2.3, together with the "exact" control $v$ obtained by Fourier expansion.

**Table 2.2: (Ex1)** – Error on the approximated solution $(\varphi_h, \lambda_h)$ of (2.20) w. r. t. $h$.

| $N_{\mathcal{T}}$ | 25 | 50 | 75 | 100 | 125 | 150 |
|---|---|---|---|---|---|---|
| $h$ ($\times 10^{-2}$) | 6.46 | 3.51 | 2.66 | 1.72 | 1.40 | 1.28 |
| $\|\varphi(\gamma, \cdot) - \varphi_h(\gamma, \cdot)\|_{\mathbf{H}}$ ($\times 10^{-1}$) | 2.15 | 1.59 | 1.31 | 1.20 | 1.09 | 1.01 |
| $\|\lambda - \lambda_h\|_{\mathbf{\Lambda}}$ ($\times 10^{-2}$) | 11.0 | 8.06 | 6.69 | 6.05 | 5.38 | 4.81 |

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

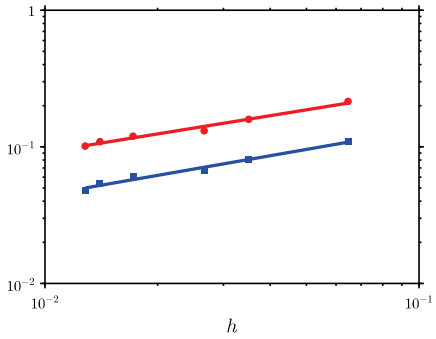



**Figure 2.1:** (**Ex1**) – Error on the approximated solution $(\varphi_h, \lambda_h)$ of (2.20) vs. $h$ – $\|\varphi(\gamma,\cdot) - \varphi_h(\gamma,\cdot)\|_{\mathbf{H}}$ (•), $\|\lambda - \lambda_h\|_{\mathbf{\Lambda}}$ (▪).

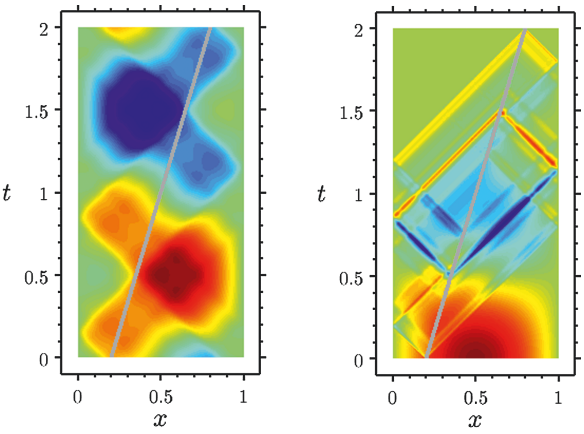



**Figure 2.2:** (**Ex1**) – Isovalues of the adjoint state $\varphi_h$ (left) and controlled state $\lambda_h$ (right) for $N_{\mathcal{T}} = 150$.

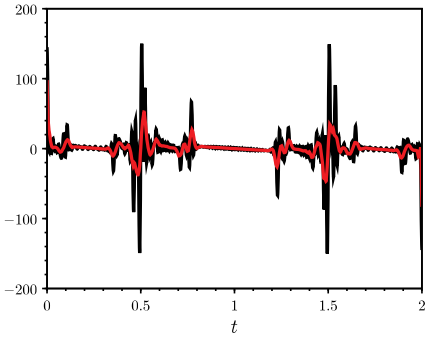



**Figure 2.3:** (**Ex1**) – Controls $v_h$ (–) and $v$ (–) for $N_{\mathcal{T}} = 150$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

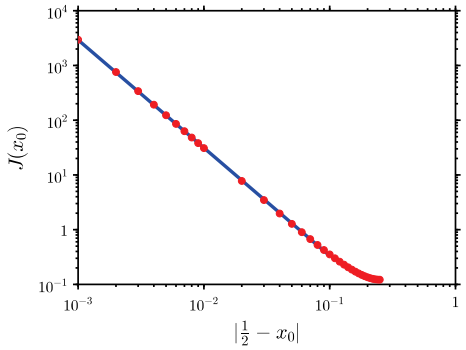



**Figure 2.4:** (**Ex2**) – $J(x_0)$ vs. $|x^\star - x_0|$ for stationary control points $x_0$.

## 2.4.2 Blowup at nonstrategic points

In the case of a stationary control point $\gamma \equiv x_0 \in \Omega$, it is well known that we have to choose a so-called strategic point (see [26]) to ensure the controllability of (2.1). A point $x_0$ is strategic if and only if $\sin(p\pi x_0) \neq 0$ for every $p \geq 1$. Moreover, a given initial datum $(y_0, y_1) \in \mathbf{V}$ can be controlled if and only if $\sin(p\pi x_0) \neq 0$ for every $p \geq 1$ such that one of the Fourier coefficients $c_p(y_0)$ and $c_p(y_1)$ is nonzero. Therefore, for fixed $(y_0, y_1) \in \mathbf{V}$, we expect the cost of control to blow up as $x_0$ gets closer to a nonstrategic location. To illustrate this property, we use the initial datum

$$y_0(x) = \sin(2\pi x), \quad y_1(x) = 0, \quad x \in \Omega, \qquad (\mathbf{Ex2}\text{–}\mathbf{y}_0)$$

and we evaluate the functional $J(x_0)$ (cf. (2.11)) for several control locations $x_0$ spread in the interval $(\frac{1}{4}, \frac{1}{2})$. With the initial datum considered, $x^\star = \frac{1}{2}$ is the unique nonstrategic point. In Figure 2.4, we display $J(x_0)$ with respect to the distance $|x^\star - x_0|$. As expected, we note that the cost of control blows up as $x_0 \to x^\star$. More precisely, we have $J(x_0) \sim_{x^\star} C_0 |x^\star - x_0|^{-1.97}$.

## 2.4.3 Optimization of the support using splines

We now focus on solving numerically problem (2.12) with a gradient-type algorithm. To do so, the control trajectories $\gamma$ considered are degree 5 splines adapted to a fixed subdivision of $[0, T]$. For any integer $N \geq 1$, we denote $S_N = (t_i)_{0 \leq i \leq N}$ the regular subdivision of $[0, T]$ in $N$ intervals. With $\kappa = T/N$, the subdivision points are $t_i = i\kappa$. In the simulations below, we use $N = 20$. We then define the set $\mathcal{S}_5$ of degree 5 splines adapted to the subdivision $S_N$. Such a spline $\gamma \in \mathcal{S}_5$ is of class $C^2([0, T])$ and is uniquely determined by the $3(N + 1)$ conditions

$$\gamma(t_i) = x_i, \quad \gamma'(t_i) = p_i, \quad \gamma''(t_i) = c_i, \quad 0 \leq i \leq N,$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

where $\mathbf{x} = (x_i)_{0\le i\le N}$, $\mathbf{p} = (p_i)_{0\le i\le N}$, and $\mathbf{c} = (c_i)_{0\le i\le N}$ represent the spline parameters. We also introduce the degree 5 polynomial basis $(P_{k,l})_{\substack{k=0,1,2\\ l=0,1}}$ on $[0,1]$ characterized by

$$P_{k,l}^{(k')}(l') = \delta_{k,k'}\,\delta_{l,l'} \quad \text{for } k,k' \in \{0,1,2\},\ l,l' \in \{0,1\}.$$

Here $P_{k,l}^{(k')}$ stands for the $k'$th derivative of $P_{k,l}$, and $\delta_{k,k'}$ is the Kronecker delta, i. e., $\delta_{k,k'} = 1$ if $k = k'$ and $\delta_{k,k'} = 0$ otherwise. For the sake of presentation, we briefly rename the parameters $(\mathbf{x},\mathbf{p},\mathbf{c}) = (\mathbf{s}^0,\mathbf{s}^1,\mathbf{s}^2)$. This allows us to decompose $\gamma$ into

$$\gamma(t) = \sum_{i=1}^{N}\sum_{k=0}^{2}\big(\mathbf{s}_{i-1}^k P_{k,0}^i(t) + \mathbf{s}_i^k P_{k,1}^i(t)\big)\mathbb{1}_{[t_{i-1},t_i]}(t) \quad \forall t\in[0,T],$$

where we have set $P_{k,l}^i(t) = \kappa^k P_{k,l}(\frac{t-t_{i-1}}{\kappa})$. With this decomposition, the optimization problem (2.12) is reduced to a finite-dimensional problem in the space of parameters, i. e.,

$$\min_{\gamma\in\mathcal{S}_5} J_{\varepsilon,\eta}(\gamma) = \min_{\mathbf{s}} \widetilde{J}_{\varepsilon,\eta}(\mathbf{s}), \quad \text{where } \mathbf{s} = (\mathbf{x},\mathbf{p},\mathbf{c}) \in \mathbb{R}^{3(N+1)}.$$

To get a descent direction for $J_{\varepsilon,\eta}$ at $\gamma\in\mathcal{S}_5$, we consider the following variational problem: find a solution $j_\gamma \in \mathcal{S}_5$ of

$$\begin{aligned}\langle j_\gamma,\overline{\gamma}\rangle_{\mathbf{H}} + \varepsilon\langle j_\gamma'',\overline{\gamma}''\rangle_{L^2(0,T)} &= \mathrm{d}J(\gamma;\overline{\gamma}) + \varepsilon\langle\gamma'',\overline{\gamma}''\rangle_{L^2(0,T)}\\ &\quad + \eta\big(L(\gamma)-\overline{L}\big)^+\mathrm{d}L(\gamma;\overline{\gamma}), \quad \forall\overline{\gamma}\in\mathcal{S}_5.\end{aligned} \tag{2.24}$$

Indeed, using Lemma 2.2.3, we can see that $\mathrm{d}J_{\varepsilon,\eta}(\gamma;j_\gamma) = \|j_\gamma\|_{\mathbf{H}}^2 + \varepsilon\|j_\gamma''\|_{L^2(0,T)}^2 \ge 0$. Problem (2.24) is solved by the finite element method using FreeFEM++. We denote by $P_\Omega$ the projection in $\Omega$. Then the gradient algorithm for solving (2.12) is given by Algorithm 2.1. We point out that a remeshing of $Q_T$ is performed at each iteration to be conform with the current trajectory $\gamma_n$. We illustrate the algorithm on two examples.

**Example 1 – Sine function**

To test Algorithm 2.1, we first use the initial datum

$$y_0(x) = 10\sin(\pi x), \quad y_1(x) = 0, \quad x\in\Omega. \tag{$\mathbf{Ex3}$–$\mathbf{y}_0$}$$

We initialize the algorithm with the trajectory $\gamma_0\in\mathcal{S}_5$ associated with the parameters

$$x_i = \frac{3}{20} + \frac{1}{5}\frac{t_i}{T}, \quad p_i = \frac{1}{5T}, \quad c_i = 0, \quad 0\le i\le N. \tag{$\mathbf{Ex3}$–$\gamma_0$}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Algorithm 2.1:** Gradient descent.

**Initialization** Choose a trajectory $\gamma_0 \in \mathcal{S}_5$ such that $0 < \gamma_0 < 1$.

**For each** $n \geq 0$ **do**

- ▷ Compute the solution $\varphi_h$ of (2.20) associated with $\gamma_n$.
- ▷ Evaluate the costs $J(\gamma_n)$ and $J_{\varepsilon,\eta}(\gamma_n)$.
- ▷ Compute the solution $j_{\gamma_n}$ of (2.24).
- ▷ Update the trajectory $\gamma_n$ by setting

$$\gamma_{n+1} = P_\Omega(\gamma_n - \rho j_{\gamma_n}) \quad \text{with } \rho > 0 \text{ fixed.}$$

**End**

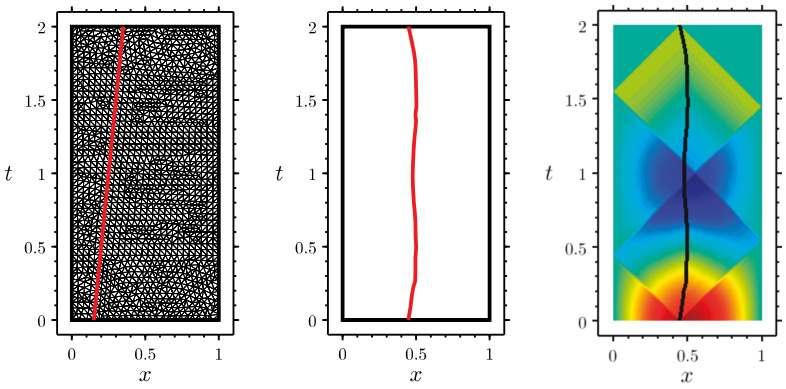



**Figure 2.5:** (**Ex3**) – Initial trajectory $\gamma_0$, optimal trajectory $\gamma^\star$, and optimal controlled state $\lambda^\star$ (from left to right). The left figure also illustrates the type of mesh used to solve (2.20).

We set $\varepsilon = 10^{-4}$, $\eta = 10^3$, $\overline{L} = 2.01$, and $\rho = 10^{-2}$. The initial trajectory $\gamma_0$, the optimal trajectory $\gamma^\star$, and the optimal controlled state $\lambda^\star$ are displayed in Figure 2.5. We observe that the optimal trajectory we get is close to a stationary control point located in $x_0 = \frac{1}{2}$, the maximum point of $\sin(\pi x)$. This is coherent with the case of controls distributed over domains $q \subset Q_T$ (see [6, Example **EX1**]).

### Example 2 – Traveling wave

To test again the similarities between the pointwise control case and the distributed control case, we now use the initial datum

$$y_0(x) = (10x-3)^2(10x-7)^2\mathbb{1}_{[0.3,0.7]}(x), \quad y_1(x) = y_0'(x), \quad x \in \Omega. \qquad (\mathbf{Ex4}\text{–}\mathbf{y}_0)$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

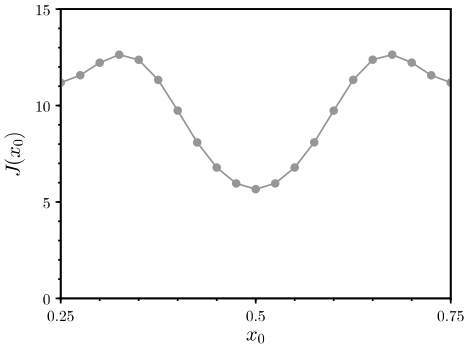



**Figure 2.6:** (**Ex4**) – $J(g_{x_0})$ vs. $x_0$.

To see whether the control trajectory is likely to "follow" the wave associated with (**Ex4**–$\mathbf{y}_0$) as it is the case in [6, Example **EX2**]), we define the trajectories

$$g_{x_0}(t) = f_{x_0}(t) + 0.15\cos\big(5\pi(t - x_0)\big) \quad \text{for any } x_0 \in \Omega.$$

Here $f_{x_0}$ is the characteristic line "$x + t = x_0$" of the wave equation. The trajectory $g_{\frac{1}{2}}$ is displayed in Figure 2.7, left. Then, for several values of $x_0$ in $\Omega$, we evaluate the functional $J(g_{x_0})$ associated with the initial datum (**Ex4**–$\mathbf{y}_0$). The results are displayed in Figure 2.6, and we can see that $J$ reaches its minimum for $x_0 = \frac{1}{2}$.

We then employ Algorithm 2.1 for two different initial trajectories $\gamma_0 \in \mathcal{S}_5$, respectively, defined by

$$x_i = g_{\frac{1}{2}}(t_i), \quad p_i = g'_{\frac{1}{2}}(t_i), \quad c_i = g''_{\frac{1}{2}}(t_i), \quad 0 \le i \le N, \qquad (\textbf{Ex4.1}\text{–}\gamma_0)$$

$$x_i = \frac{1}{4} + \frac{1}{2}\frac{t_i}{T}, \quad p_i = \frac{1}{2T}, \quad c_i = 0, \quad 0 \le i \le N. \qquad (\textbf{Ex4.2}\text{–}\gamma_0)$$

We set $\varepsilon = 10^{-4}$, $\eta = 10^3$, $\overline{L} = 4$, and $\rho = 10^{-2}$. For examples (**Ex4.1**) and (**Ex4.2**), we display the initial trajectory $\gamma_0$, the optimal trajectory $\gamma^\star$, and the optimal controlled state $\lambda^\star$ in Figures 2.7 and 2.8, respectively. In the first setup, we observe that the optimal trajectory remains close to the wave support, which is coherent with the distributed control case. In the second setup the optimal trajectory also seems to get closer to the wave support, but the convergence is very slow. This can be seen in Figure 2.9, where the evolution of the functional $J(\gamma_n)$ and the curve length $L(\gamma_n)$ are shown. The optimal costs are respectively $J(\gamma^\star) = 3.92$ for (**Ex4.1**) and $J(\gamma^\star) = 3.69$ for (**Ex4.2**). The difference is negligible compared to the initial cost $J(\gamma_0) = 37.45$ for example (**Ex4.2**).

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

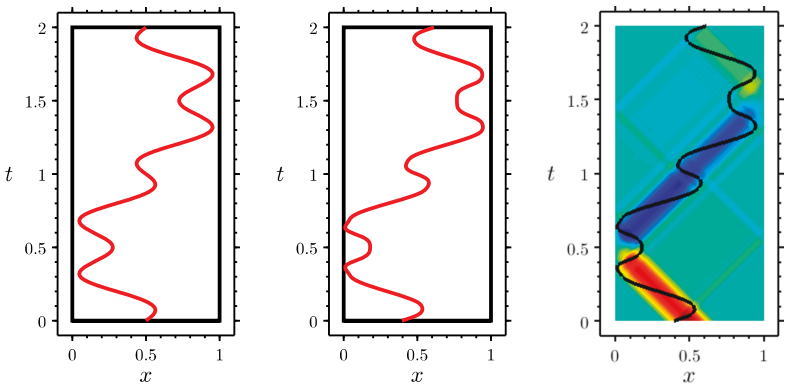



**Figure 2.7:** (**Ex4.1**) – Initial trajectory $\gamma_0$, optimal trajectory $\gamma^\star$, and optimal controlled state $\lambda^\star$ (from left to right).

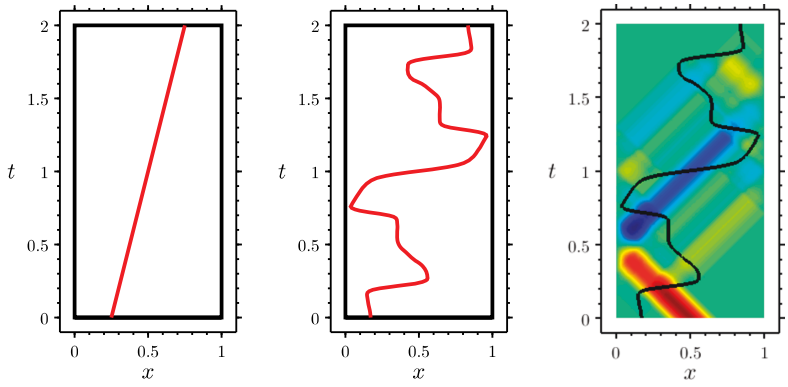



**Figure 2.8:** (**Ex4.2**) – Initial trajectory $\gamma_0$, optimal trajectory $\gamma^\star$, and optimal controlled state $\lambda^\star$ (from left to right).

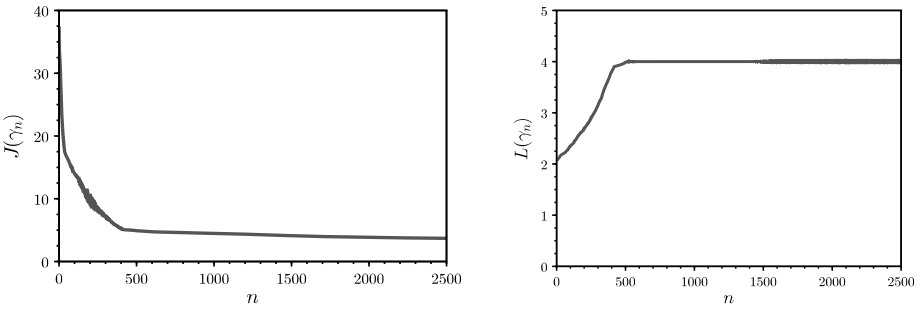



**Figure 2.9:** (**Ex4.2**) – Functional $J(\gamma_n)$ (left) and curve length $L(\gamma_n)$ (right).

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

## 2.5 Conclusion

On the basis of [9], which deals with controls distributed over noncylindrical domains, we have built a mixed formulation characterizing the HUM control acting on a moving point. The formulation involves the adjoint state and a Lagrange multiplier, which turns out to coincide with the controlled state. This approach leads to a variational formulation over a Hilbert space without distinction between the space and time variables, making it very appropriate to our moving point situation. We have shown the well-posedness of the formulation using the observability inequality proved in [8]. At a practical level the mixed formulation is discretized and solved in the finite element framework. The resolution amounts to solve a sparse symmetric system. From a numerical point of view, we have provided evidence of the convergence of the approximated control for regular initial data.

Still from a numerical perspective, for a fixed initial datum, we have considered the natural problem of optimizing the support of control. We have solved this problem with a simple gradient algorithm. For simplicity, the optimization is made over very regular trajectories. The results we get are similar to those obtained in [6], where the same problem is studied for controls distributed over noncylindrical domains, although the convergence toward the optimal trajectory seems to be generally much slower.

This work may be extended to several directions. First, as it is done in [30] for distributed controls, we could try to justify rigorously the well-posedness of the support optimization problem. In that context, it could be interesting to find the minimal regularity necessary for the control trajectories. Besides, we could try to implement other types of algorithms for solving the problem as, for instance, an algorithm based on the level-set method. Another challenge is the extension of the observability inequality to the multidimensional case, where we cannot make use of the d'Alembert formula.

## Appendix. Fourier expansion of the HUM control

In this appendix, we expand, in terms of Fourier series, the adjoint state $\varphi$ linked to the HUM control $v$ by relation (2.9), and the associated controlled state $y$. These expansions are used to evaluate the errors $\|v - v_h\|_{\mathbf{H}'}$ and $\|y - y_h\|_{\mathbf{\Lambda}}$ in Section 2.4. We can show that $\varphi$ and $y$ take the form

$$\varphi(x,t) = \sum_{p\geq 1}\left(a_p \cos(p\pi t) + \frac{b_p}{p\pi}\sin(p\pi t)\right)\sin(p\pi x), \tag{2.25}$$

$$y(x,t) = \sum_{p\geq 1} c_p(t)\sin(p\pi x). \tag{2.26}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

We set

$$\begin{cases} \xi_p^a(t) = \cos(p\pi t)\sin\big(p\pi\gamma(t)\big), \\ \xi_p^b(t) = \frac{1}{p\pi}\sin(p\pi t)\sin\big(p\pi\gamma(t)\big), \end{cases} \qquad p \geq 1.$$

Injecting (2.25) into the terms appearing in the optimality condition (2.8), we get

$$\begin{aligned} \int_0^T \varphi(\gamma(t),t)\overline{\varphi}(\gamma(t),t)\,\mathrm{d}t = {} & \sum_{p,q\geq 1} a_p\overline{a}_q \int_0^T \xi_p^a\xi_q^a + \sum_{p,q\geq 1} b_p\overline{b}_q \int_0^T \xi_p^b\xi_q^b \\ & + \sum_{p,q\geq 1} a_p\overline{b}_q \int_0^T \xi_p^a\xi_q^b + \sum_{p,q\geq 1} b_p\overline{a}_q \int_0^T \xi_p^b\xi_q^a, \end{aligned} \tag{2.27}$$

$$\begin{aligned} \int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\varphi(\gamma(t),t)\frac{\mathrm{d}}{\mathrm{d}t}\overline{\varphi}(\gamma(t),t)\,\mathrm{d}t = {} & \sum_{p,q\geq 1} a_p\overline{a}_q \int_0^T \xi_p^{a\prime}\xi_q^{a\prime} + \sum_{p,q\geq 1} b_p\overline{b}_q \int_0^T \xi_p^{b\prime}\xi_q^{b\prime} \\ & + \sum_{p,q\geq 1} a_p\overline{b}_q \int_0^T \xi_p^{a\prime}\xi_q^{b\prime} + \sum_{p,q\geq 1} b_p\overline{a}_q \int_0^T \xi_p^{b\prime}\xi_q^{a\prime}, \end{aligned} \tag{2.28}$$

$$\int_\Omega y_0\overline{\varphi}_1 = \frac{1}{2}\sum_{p\geq 1} c_p(y_0)\overline{b}_p \quad \text{and} \quad \langle y_1, \overline{\varphi}_0\rangle_{-1,1} = \frac{1}{2}\sum_{p\geq 1} c_p(y_1)\overline{a}_p, \tag{2.29}$$

where $c_p(y_0)$ and $c_p(y_1)$ are the Fourier coefficients of $y_0$ and $y_1$. Thus the optimality condition (2.8) can be rewritten

$$\left\langle \mathcal{M}_\gamma \begin{pmatrix} \{a_p\}_{p\geq 1} \\ \{b_p\}_{p\geq 1} \end{pmatrix}, \begin{pmatrix} \{\overline{a}_q\}_{q\geq 1} \\ \{\overline{b}_q\}_{q\geq 1} \end{pmatrix} \right\rangle = \left\langle \mathcal{F}_{\mathbf{y}_0}, \begin{pmatrix} \{\overline{a}_q\}_{q\geq 1} \\ \{\overline{b}_q\}_{q\geq 1} \end{pmatrix} \right\rangle, \ (\overline{a}_q, \overline{b}_q)_{q\geq 1}, \tag{2.30}$$

where the positive definite matrix $\mathcal{M}_\gamma$ and the vector $\mathcal{F}_{\mathbf{y}_0}$ are obtained from (2.27)–(2.28) and (2.29), respectively. The resolution of the infinite-dimensional system (2.30) (reduced to a finite-dimensional one by truncation) provides an approximation of the adjoint state $\varphi$ linked to the HUM control $v$ by (2.9).

Injecting (2.26) into the wave equation (2.1), we find that $c_p(t)$ satisfies

$$\begin{cases} c_p''(t) + (p\pi)^2 c_p(t) = 2v(t)\sin\big(p\pi\gamma(t)\big), \quad t > 0, \\ c_p(0) = c_p(y_0), \quad c_p'(0) = c_p(y_1). \end{cases}$$

We then have

$$c_p(t) = c_p(y_0)\cos(p\pi t) + \frac{c_p(y_1)}{p\pi}\sin(p\pi t) + \frac{2}{p\pi}\int_0^t v(s)\sin\big(p\pi\gamma(s)\big)\sin\big(p\pi(t-s)\big)\,\mathrm{d}s.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Finally, by integration by parts we deduce

$$\begin{aligned} c_p(t) = c_p(y_0)\cos(p\pi t) + \frac{c_p(y_1)}{p\pi}\sin(p\pi t) \\ + \frac{2}{p\pi}\int_0^t \varphi(\gamma(s),s)\sin\big(p\pi\gamma(s)\big)\sin\big(p\pi(t-s)\big)\,\mathrm{d}s \\ - 2\int_0^t \frac{\mathrm{d}}{\mathrm{d}s}\varphi(\gamma(s),s)\sin\big(p\pi\gamma(s)\big)\cos\big(p\pi(t-s)\big)\,\mathrm{d}s \\ + 2\int_0^t \frac{\mathrm{d}}{\mathrm{d}s}\varphi(\gamma(s),s)\cos\big(p\pi\gamma(s)\big)\gamma'(s)\sin\big(p\pi(t-s)\big)\,\mathrm{d}s. \end{aligned}$$

# Bibliography


[1] A. Agresti, D. Andreucci, and P. Loreti. Observability for the wave equation with variable support in the Dirichlet and Neumann cases. In *International Conference on Informatics in Control, Automation and Robotics*, pages 51–75. Springer, 2018.

[2] A. Bamberger, J. Jaffre, and J.-P. Yvon. Punctual control of a vibrating string: numerical analysis. *Comput. Math. Appl.*, 4(2):113–138, 1978.

[3] C. Bardos, G. Lebeau, and J. Rauch. Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary. *SIAM J. Control Optim.*, 30(5):1024–1065, 1992.

[4] M. Berggren and R. Glowinski. Controllability issues for flow-related models: a computational approach. Technical report, 1994.

[5] M. Bernadou and K. Hassan. Basis functions for general Hsieh–Clough–Tocher triangles, complete or reduced. *Int. J. Numer. Methods Eng.*, 17(5):784–789, 1981.

[6] A. Bottois, N. Cîndea, and A. Münch. Optimization of non-cylindrical domains for the exact null controllability of the 1D wave equation. *ESAIM Control Optim. Calc. Var.*, 27:13–32, 2021.

[7] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*, volume 15 of Springer Series in Computational Mathematics. Springer-Verlag, New York, 1991.

[8] C. Castro. Exact controllability of the 1-D wave equation from a moving interior point. *ESAIM Control Optim. Calc. Var.*, 19(1):301–316, 2013.

[9] C. Castro, N. Cîndea, and A. Münch. Controllability of the linear one-dimensional wave equation with inner moving forces. *SIAM J. Control Optim.*, 52(6):4027–4056, 2014.

[10] D. Chapelle and K.-J. Bathe. The inf–sup test. *Comput. Struct.*, 47(4-5):537–545, 1993.

[11] F. Chatelin. *Eigenvalues of matrices*, volume 71 of Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2012. With exercises by Mario Ahués and the author, Translated with additional material by Walter Ledermann, Revised reprint of the 1993 edition [MR1232655].

[12] P. Ciarlet. *The finite element method for elliptic problems*, volume 40 of Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. Reprint of the 1978 original [North-Holland, Amsterdam; MR0520174 (58 #25001)].

[13] N. Cîndea and A. Münch. A mixed formulation for the direct approximation of the control of minimal $L^2$-norm for linear type wave equations. *Calcolo*, 52(3):245–288, 2015.


EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[14] R. Dáger and E. Zuazua. *Wave Propagation, Observation and Control in* 1*-d Flexible Multi-Structures*, volume 50 of Mathématiques & Applications (Berlin) [Mathematics & Applications]. Springer-Verlag, Berlin, 2006.
[15] S. Ervedoza and E. Zuazua. A systematic method for building smooth controls for smooth data. *Discrete Contin. Dyn. Syst., Ser. B*, 14(4):1375–1401, 2010.
[16] R. Glowinski, J.-L. Lions, and J. He. *Exact and Approximate Controllability for Distributed Parameter Systems: A Numerical Approach*, volume 117 of Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge, 2008.
[17] F. Hecht. New development in FreeFem++. *J. Numer. Math.*, 20(3–4):251–265, 2012.
[18] A. Henrot and J. Sokołowski. A shape optimization problem for the heat equation. In *Optimal Control (Gainesville, FL, 1997)*, volume 15 of Applied Optimization, pages 204–223. Kluwer Acad. Publ., Dordrecht, 1998.
[19] A. Henrot, W. Horn, and J. Sokołowski. Domain optimization problem for stationary heat equation. *Appl. Math. Comput. Sci.*, 6(2):353–374, 1996. Shape optimization and scientific computations (Warsaw, 1994).
[20] A. Khapalov. Controllability of the wave equation with moving point control. *Appl. Math. Optim.*, 31(2):155–175, 1995.
[21] A. Khapalov. Mobile point controls versus locally distributed ones for the controllability of the semilinear parabolic equation. *SIAM J. Control Optim.*, 40(1):231–252, 2001.
[22] A. Khapalov. Observability and stabilization of the vibrating string equipped with bouncing point sensors and actuators. *Math. Methods Appl. Sci.*, 24(14):1055–1072, 2001.
[23] J. Le Rousseau, G. Lebeau, P. Terpolilli, and E. Trélat. Geometric control condition for the wave equation with a time-dependent observation domain. *Anal. PDE*, 10(4):983–1015, 2017.
[24] J.-L. Lions. *Some methods in the mathematical analysis of systems and their control*. Kexue Chubanshe (Science Press), Beijing; Gordon & Breach Science Publishers, New York, 1981.
[25] J.-L. Lions. *Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués. Tome 1*, volume 8 of Recherches en Mathématiques Appliquées [Research in Applied Mathematics]. Masson, Paris, 1988. Contrôlabilité exacte. [Exact controllability]. With appendices by E. Zuazua, C. Bardos, G. Lebeau, and J. Rauch.
[26] J.-L. Lions. Pointwise control for distributed systems. In *Control and Estimation in Distributed Parameter Systems*, pages 1–39. SIAM, 1992.
[27] J.-L. Lions and E. Magenes. *Non-Homogeneous Boundary Value Problems and Applications. Vol. I*. Springer-Verlag, New York–Heidelberg, 1972. Translated from the French by P. Kenneth, Die Grundlehren der mathematischen Wissenschaften, Band 181.
[28] A. Meyer. A simplified calculation of reduced HCT-basis functions in a finite element context. *Comput. Methods Appl. Math.*, 12(4):486–499, 2012.
[29] A. Münch. A uniformly controllable and implicit scheme for the 1-D wave equation. *M2AN Math. Model. Numer. Anal.*, 39(2):377–418, 2005.
[30] F. Periago. Optimal shape and position of the support for the internal exact control of a string. *Syst. Control Lett.*, 58(2):136–140, 2009.
[31] Á. M. Ramos, R. Glowinski, and J. Périaux. Pointwise control of the Burgers equation and related Nash equilibrium problems: computational approach. *J. Optim. Theory Appl.*, 112(3):499–516, 2002.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Martin Gugat and Michael Herty

# 3 Limits of stabilizability for a semilinear model for gas pipeline flow

**Abstract:** We present a positive and negative stabilization results for a semilinear model of gas flow in pipelines. For feedback boundary conditions, we obtain unconditional stabilization in the absence and conditional instability in the presence of the source term. We also obtain unconditional instability for the corresponding quasilinear model given by the isothermal Euler equations.

**Keywords:** stabilization, hyperbolic partial differential equations, feedback law

**MSC 2010:** 35L04, 35L65

## 3.1 Introduction

The isothermal Euler equations provide a sufficiently accurate model for the temporal and spatial evolution of gas flow in a high-pressure, intercontinental pipeline; see [3, 4, 13, 22, 25, 26]. Recent investigations have been focusing on suitable simplified models and the corresponding numerical schemes [2]. Therefore, for operational purposes of gas networks, a relevant problem is an efficient control of the underlying gas dynamics. Major physical effects for high-pressure gas pipes are the pipe wall friction and conservation of mass [22]. As a simplified model, this leads to the study of control aspects of the hyperbolic system of isothermal Euler equations (3.1). Here we state them in terms of the pressure $p = p(t,x) > 0$ and the volume flux $q = q(t,x)$. The parameters are the (constant) speed of sound $c$ and $\theta > 0$, which is proportional to the pipe wall friction factor.

$$\begin{cases} \frac{1}{(c)^2}\, p_t + q_x = 0, \\ q_t + ((1 + \frac{c^2 q^2}{p^2})p)_x = -\frac{1}{2}\theta\, c^2\, \frac{q\,|q|}{p}. \end{cases} \tag{3.1}$$

**Acknowledgement:** This work has been supported by BMBF ENets 05M18PAA, 320021702/GRK2326, 333849990/IRTG-2379, DFG 18, 19 and by DFG TRR 154, projects C03 and C05. We also would like to thank the organizers of the RICAM Seminar on Optimization 2019 for providing stimulating discussions.

**Martin Gugat,** Department Mathematik, Friedrich-Alexander Universität Erlangen-Nürnberg (FAU), Cauerstr. 11, 91058 Erlangen, Germany, e-mail: martin.gugat@fau.de
**Michael Herty,** RWTH Aachen University, Institut für Geometrie und Praktische Mathematik, Templergraben 55, 52056 Aachen, Germany, e-mail: herty@igpm.rwth-aachen.de

https://doi.org/10.1515/9783110695984-003

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

The previous set of equations is posed on $t \geq 0$ and $x \in [0, L]m$ where $L > 0$ is the length of the pipe. The initial conditions are typically prescribed as

$$p(0, x) = p_0(x) \quad \text{and} \quad q(0, x) = q_0(x). \tag{3.2}$$

Different boundary conditions have been suggested to accompany equations (3.1) and (3.2). For $p > 0$, system (3.1) forms a strictly hyperbolic system of balance laws and thus can be stated in quasilinear form using Riemann invariants. Linear feedback boundary conditions in terms of Riemann invariants have been discussed in detail in [12]. Therein, a Dirichlet feedback law is presented such that for sufficiently small feedback parameters and sufficiently short pipes, we have the exponential stability of nonconstant stationary states in the $L^2$-sense. For many practically relevant situations, a semilinear model derived from (3.1) is useful. For slowly moving gas, we have $c^2q^2/p^2 \ll 1$ for the momentum term. Hence the semilinear model studied, for example, in [17] and [19] is given by

$$\begin{cases} \frac{1}{(c)^2}\, p_t + q_x = 0, \\ q_t + p_x = -\frac{1}{2}\theta\, c^2\, \frac{q\,|q|}{p}. \end{cases} \tag{3.3}$$

This model also has constant stationary solutions $(\bar{q}, \bar{p}) = (0, \bar{p})$, where $\bar{p}$ is a real number, and we assume that $\bar{p} > 0$ according to physical considerations. On the finite space interval $[0, L]$ of length $L > 0$, we may therefore expect that the system state for any initial data $(p_0, q_0)$ as in (3.2) can be stabilized to a solution of this type by the boundary conditions

$$p_x(t, 0) = f_0\, p_t(t, 0), \quad v_x(t, L) = -\frac{1}{c}\, v_t(t, L), \tag{3.4}$$

where $f_0$ is the real feedback parameter. Furthermore,

$$v = \frac{c^2\, q}{p}$$

is the gas velocity. Note that due to the semilinear structure of equation (3.3) and the separated eigenvalues of the Jacobian of the flux, we only require one boundary condition at each end of the pipe.

The boundary conditions can also be stated in terms of $p$ and $q$. Note that we have

$$v_x(t, L) = c^2\left(\frac{q_x(t, L)}{p(t, L)} - \frac{q(t, L)p_x(t, L)}{p^2(t, L)}\right)$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

and a similar equation for $v_t$. Thus we obtain the boundary condition at $x = L$ from (3.4) in the form

$$q_x(t, L) = \frac{q(t, L)}{p(t, L)} p_x(t, L) - \frac{1}{c}\left(q_t(t, L) - \frac{q(t, L)}{p(t, L)} p_t(t, L)\right). \tag{3.5}$$

In this paper, we analyze the stability of the closed-loop system (3.3), (3.2) and boundary conditions (3.4). In particular, we study stabilization properties dependent on $\theta$. In the case $\theta = 0$, we obtain stabilization of stationary states $(0, \bar{p})$ for all positive values of $f_0 > 0$; see Theorem 3.2.1. On the other hand, for $\theta > 0$, we obtain the instability of solutions for $f_0 > \frac{1}{c}$; see Theorem 3.3.1. The analysis of instable solutions is based upon the construction of explicit analytical solutions, similar to that in [16]. To relate the results to existing stabilization results, we recall that in [12], we have shown that for the quasilinear isothermal Euler equations (3.1) and a Dirichlet–Riemann feedback law, with sufficiently small values of $f_0$ and sufficiently short pipes, we have the exponential stability. Instability results for sufficiently large values of $f_0$ for the quasilinear equations are not yet known.

## 3.2 Stability of the semilinear model without source term ($\theta = 0$)

In this section, we consider the closed-loop system (3.3), (3.2) and boundary conditions (3.4). The source term is not active, i. e., in the following section, we set

$$\theta = 0.$$

The main result is the unconditional local exponential stability for all feedback parameters $f_0 > 0$. The proof is based on the construction of a suitable Lyapunov function $E_\mu$. Such functions have been first introduced and investigated in the context of the control of shallow-water equations; see, e. g., [5, 6, 20]. Also, they have been applied in the context of gas dynamics in one-dimensional pipes and networks; see, e. g., [7–10, 14]. Further, general one-dimensional systems have been investigated using similar Lyapunov functions [21, 27, 28].

We also show that the system is finite-time stable: After the finite time $\frac{2L}{c}$, the system state remains constant. The finite-time stabilization of a single string is mentioned in [18]. The finite-time stabilization of a network of strings is studied in [15] and [1]. Finite-time control for linear evolution equation in a Hilbert space is studied in [24]. The finite-time stabilization of hyperbolic systems with zero source term over a bounded interval is studied in [23].

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Theorem 3.2.1.** *Let $\theta = 0$ and $f_0 > 0$. Let a steady state of equation* (3.3) *of the form* $(0, \bar{p})$ *be given, where we assume that $\bar{p} > 0$. Fix a terminal time $T > 0$. Let $\mu > 0$.*

*Then there exists $\varepsilon > 0$ such that for all initial datum* (3.2) $(q_0, p_0) \in H^1(0, L) \times H^1(0, L)$ *compatible with the boundary conditions* (3.4) *such that*

$$\|(q_0, p_0 - \bar{p})\|_{H^1(0,L)\times H^1(0,L)} \leq \varepsilon,$$

*the Lyapunov function $E_\mu(t)$*

$$= \frac{1}{2} \int_0^L \mathrm{e}^{-\mu x} \left( \frac{1}{c} p_x(t, x) + q_x(t, x) \right)^2 + \mathrm{e}^{\mu x} \left( \frac{1}{c} p_x(t, x) - q_x(t, x) \right)^2 dx \tag{3.6}$$

*with the solution $(q, p)$ to* (3.2), (3.3), *and* (3.4) *decays exponentially with rate $\mu$. More precisely, for all $t \in [0, T]$, we have the inequality*

$$E_\mu(t) \leq \exp(-c\,\mu t)\, E_\mu(0).$$

*Moreover, for $t > \frac{2L}{c}$, we have $p_x(t, x) = q_x(t, x) = 0$.*

*Proof.* Since system (3.3) with $\theta = 0$ is linear, its solution $(q, p)$ can be constructed using the method of characteristics. Concerning the $H^1$-compatibility of the initial conditions and the boundary conditions, the following remark is appropriate: On account of the form of the boundary conditions stated in terms of derivatives (which are not required to be continuous for $H^1$-solutions) for the $H^1$-solutions, no additional compatibility conditions are required.

At some point below in the proof, initial states in $H^1$ are approximated by initial states given by $H^2$-functions. The approximating $H^2$-functions should be chosen to satisfy $q_x(t, 0) = p_x(t, 0) = 0$ and $q_x(t, L) = p_x(t, L) = 0$. Then at $x = 0$, we have $p_x(t, 0) = 0$ and $f_0\, p_t(t, 0) = -f_0\, c^2\, q_x(t, 0) = 0$, and hence at $x = 0$, the $C^1$-compatibility conditions hold, and thus also the $H^2$-compatibility conditions. At $x = L$, we have $q_x(t, L) = 0$, $p_t(t, L) = c^2 q_x(t, L) = 0$, and $q_t(t, L) = -p_x(t, L) = 0$. This implies

$$\frac{q(t, L)}{p(t, L)}\, p_x(t, L) - \frac{1}{c} \left( q_t(t, L) - \frac{q(t, L)}{p(t, L)}\, p_t(t, L) \right) = 0,$$

and hence also at $x = L$, the $C^1$-compatibility conditions hold, and thus also the $H^2$-compatibility conditions. This approximation by $H^2$-functions is used below.

Due to the assumption on $(q_0, p_0)$, we also obtain an a priori bound on the $H^1$ norm:

$$\begin{aligned} &\|(q(t, \cdot), p(t, \cdot) - \bar{p})\|_{H^1(0,L)\times H^1(0,L)} \\ &\quad \leq C_0\, \exp(\eta\, t)\, \|(q_0, p_0 - \bar{p})\|_{H^1(0,L)\times H^1(0,L)} \end{aligned}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

for some constants $C_0 > 0$ and $\eta > 0$ independent of $T$ and initial datum. Since $\bar{p} > 0$ and $H^1$ embeds to $C^0$, we may choose $\varepsilon > 0$ sufficiently small such that $p(t, x) > 0$ for all $t \in [0, T]$ and $x \in [0, L]$. Therefore we have

$$\gamma_0 = \min_{(t,x)\in[0,T]\times[0,L]} p(t,x) > 0.$$

Assume that $\varepsilon > 0$ is chosen sufficiently small such that for all $t \in [0, T]$, we have

$$|q(t, L)| < \frac{\gamma_0}{c}. \tag{3.7}$$

Using equation (3.5), we write the boundary condition (3.4) as

$$\begin{aligned}&\frac{1}{c}p_x(t, L) - q_x(t, L)\\ &\quad = \frac{1}{c}p_x(t, L) - \frac{q(t, L)}{p(t, L)}\,p_x(t, L) + \frac{1}{c}\left(q_t(t, L) - \frac{q(t, L)}{p(t, L)}\,p_t(t, L)\right)\end{aligned}$$

for $t \in [0, T]$ almost everywhere. In fact, this is an equation in $L^2(0, T)$. Since $(p, q)$ are solutions to the differential equation (3.3) and since $\theta = 0$, for almost every $t \in [0, T]$, we have

$$\begin{aligned}\frac{1}{c}p_x(t, L) - q_x(t, L) &= -\frac{q(t, L)}{p(t, L)}\,p_x(t, L) - \frac{1}{c}\frac{q(t, L)}{p(t, L)}\,p_t(t, L)\\ &= -\frac{c\,q(t, L)}{p(t, L)}\left(\frac{1}{c}\,p_x(t, L) + \frac{1}{c^2}\,p_t(t, L)\right)\\ &= -\frac{c\,q(t, L)}{p(t, L)}\left(\frac{1}{c}\,p_x(t, L) - q_x(t, L)\right).\end{aligned}$$

Again this is an equation in $L^2(0, T)$. This implies

$$\left(1 + \frac{c\,q(t, L)}{p(t, L)}\right)\left(\frac{1}{c}\,p_x(t, L) - q_x(t, L)\right) = 0.$$

Due to (3.7), we have $\frac{c\,|q(t,L)|}{p(t,L)} \neq 1$. Thus we can divide by the continuous function $1 + \frac{c\,q(t,L)}{p(t,L)}$ that has no roots on $[0, T]$, and we obtain

$$\frac{1}{c}p_x(t, L) - q_x(t, L) = 0. \tag{3.8}$$

Choose $\mu > 0$ arbitrarily large. For initial states in $H^2(0, L) \times H^2(0, L)$ that are $H^2$-compatible with the boundary conditions (see the remarks about the compatibility conditions at the beginning of the proof), the time derivative of $E_\mu(t)$ fulfills the

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

equation

$$\begin{aligned}
\frac{d}{dt}E_\mu(t) &= \int_0^L \mathrm{e}^{-\mu x}\left(\frac{1}{c}p_x(t,x)+q_x(t,x)\right)\left(\frac{1}{c}p_{xt}(t,x)+q_{xt}(t,x)\right)\\
&\quad + \mathrm{e}^{\mu x}\left(\frac{1}{c}p_x(t,x)-q_x(t,x)\right)\left(\frac{1}{c}p_{xt}(t,x)-q_{xt}(t,x)\right)dx\\
&= \int_0^L \mathrm{e}^{-\mu x}\left(\frac{1}{c}p_x(t,x)+q_x(t,x)\right)\left(-c\,q_{xx}(t,x)-p_{xx}(t,x)\right)\\
&\quad + \mathrm{e}^{\mu x}\left(\frac{1}{c}p_x(t,x)-q_x(t,x)\right)\left(-c\,q_{xx}(t,x)+p_{xx}(t,x)\right)dx\\
&= \int_0^L -c\,\mathrm{e}^{-\mu x}\left(\frac{1}{c}p_x(t,x)+q_x(t,x)\right)\left(q_x(t,x)+\frac{1}{c}p_x(t,x)\right)_x\\
&\quad + c\,\mathrm{e}^{\mu x}\left(\frac{1}{c}p_x(t,x)-q_x(t,x)\right)\left(\frac{1}{c}p_x(t,x)-q_x(t,x)\right)_x dx\\
&= \frac{c}{2}\int_0^L -\mathrm{e}^{-\mu x}\left[\left(\frac{1}{c}p_x(t,x)+q_x(t,x)\right)^2\right]_x\\
&\quad + \mathrm{e}^{\mu x}\left[\left(\frac{1}{c}p_x(t,x)-q_x(t,x)\right)^2\right]_x dx.
\end{aligned}$$

Integration by parts and applying the boundary conditions (3.4) yield

$$\begin{aligned}
\frac{d}{dt}E_\mu(t) &= -\frac{c\,\mu}{2}\int_0^L \mathrm{e}^{-\mu x}\left(\frac{p_x(t,x)}{c}+q_x(t,x)\right)^2 + \mathrm{e}^{\mu x}\left(\frac{p_x(t,x)}{c}-q_x(t,x)\right)^2 dx\\
&\quad - \frac{c}{2}\mathrm{e}^{-\mu L}\left(\frac{1}{c}p_x(t,L)+q_x(t,L)\right)^2 + \frac{c}{2}\mathrm{e}^{\mu L}\left(\frac{1}{c}p_x(t,L)-q_x(t,L)\right)^2\\
&\quad - \frac{c}{2}\left[-\left(\frac{1}{c}p_x(t,0)+q_x(t,0)\right)^2 + \left(\frac{1}{c}p_x(t,0)-q_x(t,0)\right)^2\right]\\
&= -c\,\mu\,E(t) - 2\,c\,\mathrm{e}^{-\mu L}(q_x(t,L))^2 + 0 - \frac{c}{2}\left(-\frac{4}{c}p_x(t,0)q_x(t,0)\right)\\
&\le -c\,\mu\,E(t) + 2p_x(t,0)\,q_x(t,0) = -c\,\mu\,E(t) + 2f_0\,p_t(t,0)\,q_x(t,0)\\
&= -c\,\mu\,E(t) - 2f_0\,c^2\,(q_x(t,0))^2.
\end{aligned}$$

Since $f_0 > 0$, this implies the integral inequality

$$E_\mu(t) \le E_\mu(0) - \int_0^t c\,\mu E_\mu(\tau)\,d\tau$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

for all $t \in [0, T]$. By the density of $H^2(0, L)$ in $H^1(0, L)$ this implies that we have this inequality also for initial states in $H^1(0, L)$. Hence Grönwall's lemma yields the exponential decay of the energy $E_\mu$.

Note that the $E_\mu(t)$ is equivalent to the $L^2$-norm of $(p_x, q_x)$. In fact, we obtain a bound on this norm by the following inequalities:

$$
\begin{aligned}
&2\int_0^L \frac{1}{c^2}\left(p_x(t,x)\right)^2 + \left(q_x(t,x)\right)^2 dx \\
&\quad = \int_0^L \left(\frac{1}{c}p_x(t,x) + q_x(t,x)\right)^2 + \left(\frac{1}{c}p_x(t,x) - q_x(t,x)\right)^2 dx \\
&\quad \le \mathrm{e}^{\mu L}\int_0^L \mathrm{e}^{-\mu x}\left(\frac{1}{c}p_x(t,x) + q_x(t,x)\right)^2 + \mathrm{e}^{\mu x}\left(\frac{1}{c}p_x(t,x) - q_x(t,x)\right)^2 dx \\
&\quad = 2\,\mathrm{e}^{\mu L} E_\mu(t) \le 2\,\exp(\mu L - c\,\mu t)\,E_\mu(0) \\
&\quad = 2\exp\left(-\mu c\left(t - \frac{L}{c}\right)\right)\int_0^L \mathrm{e}^{-\mu x}\left(\frac{1}{c}p_x(0,x) + q_x(0,x)\right)^2 \\
&\qquad + \mathrm{e}^{\mu x}\left(\frac{1}{c}p_x(0,x) - q_x(0,x)\right)^2 dx \\
&\quad \le 2\exp\left(-\mu c\left(t - \frac{2L}{c}\right)\right)\int_0^L \left(\frac{1}{c}p_x(0,x) + q_x(0,x)\right)^2 \\
&\qquad + \left(\frac{1}{c}p_x(0,x) - q_x(0,x)\right)^2 dx.
\end{aligned}
$$

Since $\mu > 0$ can be chosen arbitrarily large, this implies $p_x(t, x) = q_x(t, x) = 0$ for $t > \frac{2L}{c}$. This finishes the proof of Theorem 3.2.1. □

Some remarks are appropriate. The condition on $(q_0, p_0)$ might seem restrictive, since we have to choose $\varepsilon$ sufficiently small to fulfill condition (3.7). Note that for the previous proof, instead of (3.7), we may use the following *assumption*. We assume that the flow remains subsonic for all $t \in [0, T]$:

$$\frac{|c\,q(t, L)|}{p(t, L)} < 1.$$

This condition is typically fulfilled for realistic gas transportation systems.

Further, note that equation (3.7) is an assumption on $q_0$ and since $|q(t,L)| \le C\exp(\eta t)\|q_0\|_{H^1}$, the previous result only holds for possibly very slow moving gas flow in pipes.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Note that although the finite-time stability result states that the flow remains steady after finite time, this does not imply $q = 0$, since for $\theta = 0$, all constant states of the form $(\bar{q}, \bar{p})$ are stationary states for the partial differential equation.

## 3.3 Instability of the semilinear model with source term ($\theta > 0$)

In this section, we consider the closed-loop system (3.3) for $\theta > 0$ with initial conditions (3.2) and boundary conditions (3.4). Theorem 3.3.1 states that there exist exponentially growing subsonic solutions to (3.3) for all feedback parameters $f_0$ that are sufficiently large in the sense that $f_0\, c > 1$. The explicit solutions are of a similar type as those introduced in [11].

**Theorem 3.3.1.** *Let $\theta > 0$ and $f_0 > \frac{1}{c}$. Then system* (3.3), (3.2), *and* (3.4) *is* not *asymptotically stable. There exist constant $(A, B, \Omega)$ and $\omega > 0$ such that for the initial data $(q_0, p_0) = (A\exp(\Omega x), B\exp(\Omega x))$, the corresponding solutions $(q, p)$ to equations* (3.3) *grow exponentially over time with rate $\omega$. More precisely, for $A < 0$, we have $B = -c^2 f_0 A$, $\Omega = \frac{1}{2}\frac{\theta}{c^2 f_0^2 - 1}$, and $\omega = \frac{\Omega}{f_0}$.*

*Proof.* For the analysis, we make the following ansatz for solutions $(q, p)$ to equation (3.3) such that $p_0 = p(0, x)$ and $q_0 = q(0, x)$:

$$q(t,\, x) = A\ \exp(\omega\, t)\ \exp(\Omega\, x), \tag{3.9}$$

$$p(t,\, x) = B\ \exp(\omega\, t)\ \exp(\Omega\, x). \tag{3.10}$$

Then, for $\omega \neq 0$ and $B \neq 0$, we have

$$\frac{p_x(t,x)}{p_t(t,x)} = \frac{\Omega}{\omega}, \quad v(t,x) = c^2 \frac{q(t,x)}{p(t,x)} = c^2 \frac{A}{B}.$$

Hence the boundary condition (3.4) holds, provided that

$$f_0 = \frac{\Omega}{\omega}.$$

System (3.3), (3.4) is unstable for $f_0 > 0$ if there exist real numbers $\Omega > 0$, $\omega = \frac{\Omega}{f_0} > 0$, $A$, and $B$ such that (3.3) and (3.4) is fulfilled. Due to the first equation in (3.3), we have

$$c^2\, \Omega\, A + \omega\, B = 0,$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

which yields

$$B = -c^2 \frac{\Omega}{\omega} A. \tag{3.11}$$

To obtain physically relevant solutions, we let $\Omega A < 0$ and $\omega > 0$. Then we have $B > 0$ and a positive pressure $p > 0$. The second equation in (3.3) yields

$$\omega A + \Omega B = -\frac{1}{2} \theta c^2 \frac{A}{B} |A|, \tag{3.12}$$

and together with (3.11) this leads to

$$A\left(\omega - c^2 \frac{\Omega^2}{\omega}\right) = \frac{1}{2} \theta \frac{\omega}{\Omega} |A|.$$

Equivalently, upon multiplication by $\frac{\Omega}{\omega}$ and since $\Omega A < 0$,

$$\frac{1}{2} \theta |A| = \Omega A \left(1 - c^2 \frac{\Omega^2}{\omega^2}\right) = -|\Omega| \, |A| \left(1 - c^2 \frac{\Omega^2}{\omega^2}\right).$$

Reformulation of the previous equality yields

$$\frac{\Omega^2}{\omega^2} = \frac{1}{c^2} \left(1 + \frac{1}{2} \frac{\theta}{|\Omega|}\right). \tag{3.13}$$

Since $\omega = \frac{\Omega}{f_0}$, we obtain

$$f_0^2 = \frac{\Omega^2}{\omega^2} > \frac{1}{c^2}.$$

Hence, for any $f_0 > \frac{1}{c}$, we define

$$\Omega = \frac{1}{2} \frac{\theta}{c^2 f_0^2 - 1} > 0 \tag{3.14}$$

and

$$\omega = \frac{\Omega}{f_0} > 0.$$

We choose a real number $A < 0$. Define $B$ by

$$B = -c^2 f_0 A > 0.$$

Then (3.11) and (3.12) hold.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

The solution (3.9)–(3.10) grows exponentially in time with rate $\omega > 0$ and satisfies (3.13) and $A\Omega < 0$. Due to the exponential growth of the state $(q, p)$, we have proved Theorem 3.3.1. □

Theorem 3.3.1 shows that the effect of the source term changes the behavior of the system dramatically also if it is arbitrarily small: For all $\theta > 0$, if the feedback parameter is greater than $\frac{1}{c}$, then the system is exponentially unstable! Note that we can choose $|A|$ arbitrarily small, and in this way, we can also make $|B|$ arbitrarily small. This implies that starting from an arbitrarily small initial state, we can obtain the exponential growth of the state in time.

## 3.4 Instability of the quasilinear model with source term ($\theta > 0$)

The instability observed for the semilinear model is also present in the quasilinear model (3.1) even for all $f_0 > 0$. A similar construction using the boundary conditions (3.4) can be obtained. We have the following result:

**Theorem 3.4.1.** *Let $\theta > 0$ and $f_0 > 0$. Then system* (3.1), (3.2), *and* (3.4) *is* not *asymptotically stable. There exist constant $(A, B, \Omega)$ and $\omega > 0$ such that for initial data $(q_0, p_0) = (A\exp(\Omega x), B\exp(\Omega x))$, the corresponding solutions $(q, p)$ to equations* (3.1) *grow exponentially over time. More precisely, for $A < 0$, we have $B = -c^2 f_0 A$, $\Omega = \frac{1}{2}\frac{\theta}{c^2 f_0^2}$, and $\omega = \frac{\Omega}{f_0}$.*

*Proof.* As in the proof of Theorem 3.3.1, we consider solutions of the type

$$p(t,x) = B\exp(\omega t)\exp(\Omega x), \quad q(t,x) = A\exp(\omega t)\exp(\Omega x)$$

and obtain the following relations for $\omega \neq 0$ and $B \neq 0$:

$$\frac{p_x(t,0)}{p_t(t,x)} = \frac{\Omega}{\omega}, \quad v(t,x) = c^2\frac{A}{B}.$$

Therefore the boundary conditions (3.4) imply

$$f_0 = \frac{\Omega}{\omega}.$$

The conservation of mass yields

$$B = -c^2 A\frac{\Omega}{\omega} = -c^2 A f_0.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

The momentum equation (3.1) yields for $(q, p)$ as above

$$\omega A + \Omega B + \frac{c^2 \Omega A^2}{B} = -\frac{1}{2}\theta c^2 \frac{A|A|}{B}. \tag{3.15}$$

Now we assume that $f_0 > 0$ and set $A = -1$. Using the value of $A$ and the relation $f_0 = \frac{\Omega}{\omega}$, we obtain

$$-\frac{\Omega}{f_0} + \Omega f_0 c^2 + \frac{c^2 \Omega}{c^2 f_0} = \frac{1}{2}\theta \frac{c^2}{f_0 c^2}.$$

This yields the positive value for $\Omega$ as

$$\begin{aligned} \Omega &= \frac{\theta}{2 f_0^2 c^2}, \\ \omega &= \frac{\Omega}{f_0} > 0, \\ B &= f_0 c^2 > 0. \end{aligned}$$

Note that if $A$ and $B$ solve (3.15), then for all $\lambda > 0$, $\lambda A$ and $\lambda B$ also solve (3.15). This finishes the proof of Theorem 3.4.1. □

Some remarks on the results of Theorems 3.3.1 and 3.4.1 are in order.

Again, we have obtained arbitrarily small initial states that generate exponentially growing solutions.

Compared to the result of Theorem 3.3.1, we have instability for *all* values $f_0 > 0$. Hence the instability properties change when considering the additional momentum flux $c^2 \frac{q^2}{p}$.

In both results (Theorems 3.3.1 and 3.4.1) the gas velocity is negative ($A < 0, B > 0$), so that the gas flows from $x = L$ in direction $x = 0$. Therefore it is reasonable that the pressure increases along the pipe. Hence the constructed solutions in both cases resemble physical intuition.

In both results, we observe that considering the limit as $\theta \to 0$, we obtain constant states

$$q(t, x) = A, \quad p(t, x) = B.$$

Hence the instability is strictly due to the presence of the source term (and the boundary conditions). Changing the later to, e. g., Dirichlet-type conditions will lead to different results as shown, e. g., in [9, 12].

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

## 3.5 Conclusions

In this paper, we have shown that a source term may change the stability properties of a system with a nonlinear feedback law. A semilinear model without source term is exponentially stable sufficiently close to an equilibrium state. However, the corresponding system with source term allows for solutions that grow exponentially. More precisely, while the system without source term is stable for all feedback parameters $f_0 > 0$, we show that including the source term yields exponentially growing in time solutions for any $f_0 > \frac{1}{c}$. The later result is established by construction of explicit analytical solutions similarly to [11, 16]. Also, the quasi-linear model with source term shows the exponential instable behavior even for any $f_0 > 0$.

## Bibliography

[1] F. Alabau-Boussouira, V. Perrollaz, and L. Rosier. Finite-time stabilization of a network of strings, *Math. Control Relat. Fields*, 5:721–742, 2015.
[2] M. K. Banda and M. Herty, Numerical discretization of stabilization problems with boundary controls for systems of hyperbolic conservation laws. *Math. Control Relat. Fields*, 3:121–142, 2013.
[3] M. K. Banda, M. Herty, and A. Klar. Coupling conditions for gas networks governed by the isothermal Euler equations. *Netw. Heterog. Media*, 1:295–314, 2006.
[4] M. K. Banda, M. Herty, and A. Klar/ Gas flow in pipeline networks. *Netw. Heterog. Media*, 1:41–56, 2006.
[5] G. Bastin, J.-M. Coron, and B. d'Andréa-Novel. On Lyapunov stability of linearised Saint-Venant equations for a sloping channel. *Netw. Heterog. Media*, 4:177–187, 2009.
[6] J.-M. Coron, B. d'Andréa-Novel, and G. Bastin. A strict Lyapunov function for boundary control of hyperbolic systems of conservation laws. *IEEE Trans. Autom. Control*, 52:2–11, 2007.
[7] M. Dick, M. Gugat, and G. Leugering. A strict $H^1$-Lyapunov function and feedback stabilization for the isothermal Euler equations with friction. *Numer. Algebra Control Optim.*, 1:225–244, 2011.
[8] M. Gugat. Optimal nodal control of networked hyperbolic systems: evaluation of derivatives. *Adv. Model. Optim.*, 7:9–37, 2005.
[9] M. Gugat. Exponential stabilization of the wave equation by Dirichlet integral feedback. *SIAM J. Control Optim.*, 53:526–546, 2015.
[10] M. Gugat, M. Dick, and G. Leugering. Gas flow in fan-shaped networks: classical solutions and feedback stabilization. *SIAM J. Control Optim.*, 49:2101–2117, 2011.
[11] M. Gugat and S. Gerster. On the limits of stabilizability for networks of strings. *Syst. Control Lett.*, 131:104494, 2019.
[12] M. Gugat and M. Herty. Existence of classical solutions and feedback stabilization for the flow in gas networks. *ESAIM Control Optim. Calc. Var.*, 17:28–51, 2011.
[13] M. Gugat, M. Herty, and V. Schleper. Flow control in gas networks: exact controllability to a given demand. *Math. Methods Appl. Sci.*, 34:745–757, 2011.
[14] M. Gugat, G. Leugering, S. Tamasoiu. and K. Wang. $H^2$-Stabilization of the isothermal Euler equations: a Lyapunov function approach. *Chin. Ann. Math., Ser. B*, 33:479–500, 2012.
[15] M. Gugat and M. Sigalotti. Stars of vibrating strings: switching boundary feedback stabilization. *Netw. Heterog. Media*, 5:299–314, 2010.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[16] M. Gugat and S. Ulbrich. The isothermal Euler equations for ideal gas with source term: product solutions, flow reversal and no blow up. *J. Math. Anal. Appl.*, 454:439–452, 2017.
[17] F. Hante, G. Leugering, A. Martin, L. Schewe, and M. Schmidt. Challenges in optimal control problems for gas and fluid flow in networks of pipes and canals: from modeling to industrial applications. In *Industrial Mathematics and Complex Systems: Emerging Mathematical Models, Methods and Algorithms*, pages 77–122. Springer Singapore, 2017.
[18] V. Komornik. Rapid boundary stabilization of the wave equation. *SIAM J. Control Optim.*, 29:197–208, 1991.
[19] G. Leugering, A. Martin, M. Schmidt, and M. Sirvent. Nonoverlapping domain decomposition for optimal control problems governed by semilinear models for gas flow in networks. *Control Cybern.*, 46:191–225, 2017.
[20] G. Leugering and E. J. P. G. Schmidt. On the modelling and stabilization of flows in networks of open canals. *SIAM J. Control Optim.* 41:164–180, 2002.
[21] T. Li, B. Rao, and Z. Wang. Exact boundary controllability and observability for first order quasilinear hyperbolic systems with a kind of nonlocal boundary conditions. *Discrete Contin. Dyn. Syst.*, 28:243–257, 2010.
[22] A. Osiadacz. *Simulation and Analysis of Gas Networks*. Gulf Publishing Company, Houston, TX, 1987.
[23] V. Perrollaz and L. Rosier. Finite-time stabilization of hyperbolic systems over a bounded interval. In *1st IFAC Workshop on Control of Systems Governed by Partial Differential Equations*, September 25–27, Paris, France, 2013.
[24] A. Polyakov, J.-M. Coron, and L. Rosier. On homogeneous finite-time control for linear evolution equation in Hilbert space. *IEEE Trans. Autom. Control*, 63(9):3143–3150, 2018.
[25] M. Schmidt, M. C. Steinbach, and B. Willert. High detail stationary optimization models for gas networks – part I: model components. *Optim. Eng.*, 16:131–164, 2015.
[26] M. C. Steinbach. On PDE solution in transient optimization of gas networks. *J. Comput. Appl. Math.*, 203:345–361, 2007.
[27] K. Wang. Exact boundary controllability of nodal profile for 1-D quasilinear wave equations. *Front. Math. China*, 6:545–555, 2011.
[28] K. Wang. Global exact boundary controllability for 1-D quasilinear wave equations. *Math. Methods Appl. Sci.*, 34:315–324, 2011.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Luis Almeida, Jesús Bellver Arnau, Michel Duprez, and Yannick Privat

# 4 Minimal cost-time strategies for mosquito population replacement

**Abstract:** Vector control plays a central role in the fight against vector-borne diseases and, in particular, arboviruses. The use of the endosymbiotic bacterium *Wolbachia* has proven to be effective in preventing the transmission of some of these viruses between mosquitoes and humans, making it a promising control tool. The population replacement technique we consider consists in replacing the wild population by a population carrying the aforementioned bacterium, thereby preventing outbreaks of the associated vector-borne diseases. In this work, we consider a two-species model incorporating both Wolbachia infected and wild mosquitoes. Our system can be controlled thanks to a term representing an artificial introduction of Wolbachia-infected mosquitoes. Under the assumption that the birth rate of mosquitoes is high, we may reduce the model to a simpler one regarding the proportion of infected mosquitoes. We investigate minimal cost-time strategies to achieve a population replacement both analytically and numerically for the simplified 1D model and only numerically for the full 2D system.

**Keywords:** minimal time, optimal control, Wolbachia, ordinary differential systems, epidemic vector control

**MSC 2010:** 49K15, 92B05, 49M05

## 4.1 Introduction

Arboviruses are a major threat for human health throughout the world, being responsible for diseases such as Dengue, Zika, Chikungunya, or Yellow fever [7, 22]. This has

**Acknowledgement:** This research was supported by the Project "Analysis and simulation of optimal shapes – application to life science" of the Paris City Hall. This program has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 754362.

**Luis Almeida, Jesús Bellver Arnau,** Sorbonne Université, CNRS, Université de Paris, Inria, Laboratoire J.-L. Lions, 75005 Paris, France, e-mails: luis.almeida@sorbonne-universite.fr, bellver@ljll.math.upmc.fr
**Michel Duprez,** Inria, équipe MIMESIS, Université de Strasbourg, Icube, CNRS UMR 7357, Strasbourg, France, e-mail: michel.duprez@inria.fr
**Yannick Privat,** IRMA, Université de Strasbourg, CNRS UMR 7501, Inria, 7 rue René Descartes, 67084 Strasbourg, France; and Institut Universitaire de France, e-mail: yannick.privat@unistra.fr

https://doi.org/10.1515/9783110695984-004

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

led to the development of increasingly sophisticated techniques to fight against these viruses, especially, techniques targeting the vector transmitting the diseases, i. e., the mosquito [3, 4, 10].

Recently, there has been increasing interest in using the well-known bacterium Wolbachia, which lives inside insect cells [9], as a tool for carrying out this vector-targeted control [6, 10, 11, 15–17].

There is a strategy named "Incompatible Insect Technique" [19] for using Wolbachia-infected male mosquito releases to control mosquito populations. It is quite similar to the more widespread sterile insect technique (they are sometimes used together). It is based on a phenomenon called cytoplasmic incompatibility (CI) [12, 18], which produces cross sterility between Wolbachia-infected males and uninfected females.

This paper is concerned with another way of using Wolbachia-infected mosquitoes to control vector populations. It takes advantage not only of CI, but also of the vertical transmission of Wolbachia in mosquitoes (it is transmitted from the mother to its offspring) and of the fact that mosquitoes carrying this bacterium show a significant reduction in their vectorial capacity for several arboviroses [13, 14, 20, 21]. Together, these properties make population replacement an interesting tool to fight against these diseases. Thus this second technique consists in releasing both males and females to achieve a replacement (at least partial) of the original mosquito population (which were efficient vectors for the disease) by one of Wolbachia-infected mosquitoes (which will not be good vectors for it).

In this work, we focus on taking advantage of mathematical modeling to find ways to minimize the amount of mosquitoes needed to ensure an effective population replacement. We will first fix the time horizon over which we will be able to act and consider a cost functional modeling the population replacement strategy. Next, a convex combination of the time horizon and the previous cost functional will be considered to get an optimized cost-time strategy. To investigate this issue, let us introduce a model for two interacting mosquito populations, a Wolbachia-free population $n_1$ and a Wolbachia carrying one, $n_2$. The resulting system, introduced in [2], reads

$$\begin{cases} \frac{dn_1(t)}{dt} = b_1 n_1(t)(1 - s_h \frac{n_2(t)}{n_1(t)+n_2(t)})(1 - \frac{n_1(t)+n_2(t)}{K}) - d_1 n_1(t), \\ \frac{dn_2(t)}{dt} = b_2 n_2(t)(1 - \frac{n_1(t)+n_2(t)}{K}) - d_2 n_2(t) + u(t), \quad t > 0, \\ n_1(0) = n_1^0, \quad n_2(0) = n_2^0, \end{cases} \tag{4.1}$$

where $u$ plays the role of a control function that will be made precise in the sequel.

The aim is to achieve a population replacement meaning that starting from the equilibrium $(n_1^*, 0) = (K(1 - \frac{d_1}{b_1}), 0)$, we want to reach the equilibrium $(0, n_2^*) = (0, K(1 - \frac{d_2}{b_2}))$. In the system the positive constant $K$ represents the carrying capacity of the environment (concerning the sum of the two populations $n_1(t) + n_2(t)$), $d_i$ and $b_i$, $i = 1, 2$, denote respectively the death and birth rates of the mosquitoes, and $s_h$ is the CI rate.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

To act on the system, we will use the control function appearing in the second equation, representing the rate at which Wolbachia-infected mosquitoes are introduced into the population.

We will impose the following biological constraint on the control function $u$: the rate at which we can instantaneously release the mosquitoes will be bounded above by $M$. Another natural biological constraint is to limit the total amount of mosquitoes we release up to time $T$, as done in [1] and [2]. In this paper, we take a different approach, looking actually at minimizing the total number of mosquitoes used. We thus introduce the space of admissible controls

$$\mathcal{U}_{T,M} := \{u \in L^\infty([0,T]), 0 \le u \le M \text{ a. e. in } (0,T)\}. \tag{4.2}$$

Before going further, to simplify the analytical study of the system, we will work on a reduction of the problem already used in [2]. It is shown there that, under the hypothesis of having a high birth rate, i. e., considering $b_1 = \frac{b_1^0}{\varepsilon}$, $b_2 = \frac{b_2^0}{\varepsilon}$ and taking the limit as $\varepsilon \to 0$, the proportion $\frac{n_2}{n_1+n_2}$ of Wolbachia-infected mosquitoes in the population uniformly converges to $p$, the solution of the simple scalar ODE

$$\begin{cases} \frac{dp}{dt}(t) = f(p(t)) + u(t)g(p(t)), \quad t > 0, \\ p(0) = 0, \end{cases} \tag{4.3}$$

where

$$f(p) = p(1-p)\frac{d_1 b_2^0 - d_2 b_1^0(1-s_h p)}{b_1^0(1-p)(1-s_h p) + b_2^0 p} \quad \text{and} \quad g(p) = \frac{1}{K}\frac{b_1^0(1-p)(1-s_h p)}{b_1^0(1-p)(1-s_h p) + b_2^0 p}.$$

We remark that in the absence of a control function, the mosquito proportion equation simplifies into $\frac{dp}{dt} = f(p)$. This system is bistable, with two stable equilibria at $p = 0$ and $p = 1$ and one unstable equilibrium at $p = \theta = \frac{1}{s_h}(1 - \frac{d_1 b_2^0}{d_2 b_1^0})$, the only root of $f$ strictly between 0 and 1 assuming that

$$1 - s_h < \frac{d_1 b_2^0}{d_2 b_1^0} < 1. \tag{4.4}$$

In what follows, it will be useful to notice that the derivative of the function $f/g$ has a unique zero $p^*$ in $(0,\theta)$,

$$p^* = \frac{1}{s_h}\left(1 - \sqrt{\frac{d_1 b_2^0}{d_2 b_1^0}}\right). \tag{4.5}$$

In [2], for the aforementioned system of two equations, the problem

$$\inf_{u \in \mathcal{U}_{T,C,M}} J(u) \quad \text{with } J(u) = \frac{1}{2}n_1(T)^2 + \frac{1}{2}\left[(n_2^* - n_2(T))_+\right]^2 \tag{4.6}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

with $\mathcal{U}_{T,C,M} = \{u \in \mathcal{U}_{T,M}, \int_0^T u(t)dt \leq C\}$ is considered. Then, denoting by $J^\varepsilon(u)$ the criterion $J(u)$ where the birth rates $b_1$ and $b_2$ have been respectively replaced by $\frac{b_1^0}{\varepsilon}$ and $\frac{b_2^0}{\varepsilon}$ with $\varepsilon > 0$, a Γ-convergence-type result is proven. More precisely, any solution of the reduced problem is close to the solutions of the original problem for the weak-star topology of $L^\infty(0,T)$, and, moreover,

$$\lim_{\varepsilon \to 0} \inf_{u \in \mathcal{U}_{T,C,M}} J^\varepsilon(u) = \inf_{u \in \mathcal{U}_{T,C,M}} J^0(u),$$

where

$$J^0(u) = \lim_{\varepsilon \to 0} J^\varepsilon(u) = K\big(1 - p(T)\big)^2, \tag{4.7}$$

and $p$ is the solution of (4.3) associated with the chosen control function $u$. The arguments exposed in [2] can be adapted to our problem, allowing us to investigate the minimization of $J^0$ given by (4.7), which is easier to study both analytically and numerically, instead of the full problem (4.6), since the solutions of both problems will be close in the sense of the Γ-convergence.

In accordance with the stability considerations above concerning system (4.3) without control, we will impose as final state constraint $p(T) = \theta$, since once we are above this state, the system will evolve by itself (with no need to control it any longer) to $p = 1$, the state of total invasion. By analogy with the two-equation system, $\theta$ represents the threshold of the basin of attraction of the equilibrium $(0, n_2^*)$. Since our goal in this work is minimizing the cost of our action on the system, we will address the issue of minimizing the total number mosquitoes released, in other words, the integral over time of the rate at which mosquitoes are being released, namely

$$J(u) = \int_0^T u(s)\, ds. \tag{4.8}$$

## 4.2 Optimal control with a finite horizon of time

We first consider an optimal control problem where the time window $[0, T]$ in which we are going to act on the system is fixed. This leads us to deal with the optimal control problem

$$\begin{cases} \inf\limits_{u \in \mathcal{U}_{T,M}} J(u), \\ p' = f(p) + ug(p), \quad p(0) = 0, \quad p(T) = \theta, \end{cases} \tag{$\mathcal{P}_{T,M}$}$$

where $J(u)$ is defined by (4.8), and $\mathcal{U}_{T,M}$ is given by (4.2).

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Theorem 4.2.1.** *Let us introduce*

$$m^* := \max_{p\in[0,\theta]}\left(-\frac{f(p)}{g(p)}\right) \quad \text{and} \quad T^* = \int_0^\theta \frac{dv}{f(v)+Mg(v)} \quad \text{for } M>0. \tag{4.9}$$

*Let us assume that $M > m^*$, $T \geq T^*$, and that* (4.4) *is true. Then there exists a bang-bang control $u^* \in \mathcal{U}_{T,M}$ solving problem $(\mathcal{P}_{T,M})$. Furthermore, every function $u^*$ defined by $u^*_\xi = M\mathbb{1}_{(0+\xi,T^*+\xi)}$, with $\xi \in [0, T-T^*]$ solves the problem, and $J(u^*_\xi) = MT^*$.*

*Proof.* Observe first that $T^*$ is constructed to be the exact time such that $p(T) = \theta$ whenever we take the "maximal" control equal to $M$ on $(0, T^*)$. For this reason, if $T = T^*$, then the set of admissible controls reduces to a singleton, and we will assume from now on that $T > T^*$.

For readability, the proof of the existence of solutions is treated separately in Appendix A. We focus here on deriving and exploiting the necessary optimality conditions. To this aim, let $u^*$ be a solution of problem $(\mathcal{P}_{T,M})$. To apply the Pontryagin maximum principle (PMP), let us introduce $U = [0, M]$ and the Hamiltonian $\mathcal{H}$ of the system given by

$$\mathcal{H} : \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R} \times \{0,-1\} \times U \ni (t,p,q,q^0,u) \mapsto q\big(f(p)+ug(p)\big) + q^0 u,$$

where $q$ satisfies

$$q'(t) = -\frac{\partial \mathcal{H}}{\partial p} = -q\big(f'(p)+ug'(p)\big), \quad t \in (0,T),$$

so that $q(t) = q(0)e^{-\int_0^t f'(p(s))+u(s)g'(p(s))ds}$, and thus $q(t)$ has a constant sign.

The instantaneous maximization condition reads

$$u^*(t) \in \arg\max_{v\in U} \mathcal{H}(t,p,q,q^0,v) = \arg\max_{v\in U}(qg(p)+q^0)v. \tag{4.10}$$

This condition implies that $q$ is positive in $(0, T)$. Indeed, assume by contradiction that $q \leq 0$ in $(0, T)$. Since $q$ has a constant sign, since $q^0 \in \{0, 1\}$, and since the pair $(q, q^0)$ is nontrivial according to the Pontryagin maximum principle, there are two possibilities: either $q^0 = -1$ and $q \leq 0$ in $(0, T)$ or $q^0 = 0$ and $q < 0$ in $(0, T)$. In both cases the optimality condition yields that $u^*(t) = 0$ for all $t \in [0, T]$, which is in contradiction with the condition $p(T) = \theta$. We thus get that $q(0) > 0$ and $q(t) > 0$ in $(0, T)$.

Now let us show that $q^0 = -1$. To this aim, let us assume by contradiction that $q^0 = 0$. Hence the optimality condition reads $u^*(t) \in \arg\max_{v\in U} qg(p)v$, and since both $q$ and $g \circ p$ are positive in $[0, T]$, we have $u^*(t) = M\mathbb{1}_{[0,T]}$. But this control function satisfies the final state constraint $p(T) = \theta$ if and only if $T = T^*$. We have thus reached a contradiction if $T > T^*$. Therefore it follows that $q^0 = -1$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Let us introduce the function $w$ given by $w(t) = q(t)g(p(t))$. The maximization condition (4.10) yields

$$\begin{cases} w(t) \leq 1 \text{ on } \{u^* = 0\}, \\ w(t) = 1 \text{ on } \{0 < u^* < M\}, \\ w(t) \geq 1 \text{ on } \{u^* = M\}. \end{cases}$$

Let us finally prove that any solution is bang-bang. The following approach is mainly inspired from [2, proof of Lemma 7]. For readability, we recall the main steps and refer to this reference for further details.

Note that

$$\begin{aligned} w'(t) &= q'(t)g(p(t)) + q(t)g'(p(t))p'(t) \\ &= -q(t)(f'(p(t)) + u(t)g'(p(t)))g(p(t)) \\ &\quad + q(t)g'(p(t))(f(p(t)) + u(t)g(p(t))) \\ &= q(t)(-f'(p(t))g(p(t)) + f(p(t))g'(p(t))) \\ &= q(t)g(p(t))^2\left(-\frac{f}{g}\right)'(p(t)). \end{aligned}$$

Looking at the monotonicity of $p \mapsto -f(p)/g(p)$, we deduce that $w$ is increasing if $0 < p(t) < p^*$ according to (4.4) and decreasing if $p^* < p(t) < \theta$, where $p^*$ is defined by (4.5). Recall that $0 < p^* < \theta$ according to (4.4).

To prove that $u^*$ is bang-bang, we show that $w$ cannot be constant on a measurable set of positive measure. By contradiction, assume that $w$ is constant on a measurable set $I$ of positive measure. Necessarily, we also have $(-f/g)'(p(t)) = 0$ on $I$, which implies that $p(t) = p^*$ on $I$. This implies that for a. e. $t \in I$, $p(t)$ is constant. To get that, we must have $p' = 0$ on that set.[1] At this step, we have

$$\{0 < u^* < M\} \subset \{t \in (0,T) \mid p(t) = p^*\} \subset \{t \in (0,T) \mid u^*(t) = -f(p^*)/g(p^*)\}.$$

Since $M > m^*$, we have $-f(p^*)/g(p^*) \in (0,M)$, which shows that the converse inclusion is true, and therefore

$$\{0 < u^* < M\} = \{t \in (0,T) \mid u^*(t) = -f(p^*)/g(p^*)\}.$$

Using that $\{w = 1\} \subset \{p = p^*\}$, we get

$$\{0 < u^* < M\} = \{w = 1\}, \quad \{u^* = M\} = \{w > 1\}, \quad \{u^* = 0\} = \{w < 1\},$$

and in particular, $\{u^* = M\}$ and $\{u^* = 0\}$ are open sets.

**1** Recall that if a function $F$ in $H^1(0,T)$ is constant on a measurable subset of positive Lebesgue measure, then its derivative is equal to 0 a. e. on $I$; see, e. g., [8, Lemma 3.1.8].

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Let $I$ be a maximal interval on which $p$ is equal to $p^*$. On $I$, we have $u^*(t) = -f(p^*)/g(p^*)$. If $u = 0$ at the end of the interval, then $p$ must decrease since $p^* < \theta$. Therefore $(-f/g)'(p(t)) > 0$, so that $w$ increases. This is in contradiction with the necessary optimality conditions (on $\{u^* = 0\}$, we have $w \leq 1$). If $u = M$ at the end of the interval, then $p$ increases, and $w$ decreases, leading again to a contradiction. Hence we must have $|I| = 0$ or $I = [0, T]$. Since $p(t) = p^*$ on $I$, $p(0) = 0$, and $p(T) = \theta$, this is impossible.

Therefore $u^*$ is equal to 0 or $M$ almost everywhere, meaning that it is bang-bang.

Finally, let us prove that the set $\{u^* = M\}$ is a single interval. We argue by contradiction. Using the facts that the solution is bang-bang, $M > m^*$, and $\{u^* = M\}$ is open, if $\{u^* = M\}$ is not one open interval, then there exists $(t_1, t_2)$ on which $u^* = 0$ and $p' < 0$. Since the final state is fixed, $p(T) = \theta$, there necessarily exists $t_3 > t_2$ such that $p(t_3) = p(t_1)$. Let us define $\tilde{u}$ as

$$\tilde{u}(t) = \begin{cases} 0, & t \in (0, t_3 - t_1), \\ u^*(t - t_3 + t_1), & t \in (t_3 - t_1, t_3), \\ u^*(t), & t \in (t_3, T). \end{cases}$$

We can easily check that since $u^* \in \mathcal{U}_{T,M}$, we have $\tilde{u} \in \mathcal{U}_{T,M}$, and if $p$ defined as the solution to

$$\begin{cases} p'(t) = f(p(t)) + u^*(t)g(p(t)), & t \in [0, T], \\ p(0) = 0 \end{cases}$$

satisfies $p(T) = \theta$, then the function $\tilde{p}$ defined as the solution to

$$\begin{cases} \tilde{p}'(t) = f(\tilde{p}(t)) + \tilde{u}(t)g(\tilde{p}(t)), & t \in [0, T], \\ \tilde{p}(0) = 0 \end{cases}$$

satisfies $\tilde{p}(T) = \theta$ and is moreover nondecreasing. Now performing a direct comparison between the cost of both controls, we obtain

$$J(u^*) - J(\tilde{u}) = \int_0^T u^*(t)dt - \int_0^T \tilde{u}(t)dt = \int_{t_2}^{t_3} u^*(t)\,dt > 0,$$

which contradicts the optimality of $u^*$. Therefore $\{u^* = M\}$ is a single interval, and it follows that

$$\{u^* = 0\} = \{p \in \{0, \theta\}\} = \{p = 0\} \cup \{p = \theta\}.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

We conclude that all the mosquitoes are released in a single interval such that $|\{u^* = M\}| = T^*$, and thus the set $\{u^* = 0\}$ splits into two intervals satisfying $|\{u^* = 0\}| = |\{p = 0\}| + |\{p = \theta\}| = T - T^*$.

Since for all $\xi \in [0, T - T^*]$, $u^*_\xi = M\mathbb{1}_{(0+\xi, T^*+\xi)}$ satisfies this property, we get that $J(u^*_\xi) = MT^*$ and there are infinitely many solutions to problem $(\mathcal{P}_{T,M})$. □

## 4.3 Optimal control with a free horizon of time

In the previous section, we obtained infinitely many solutions for problem $(\mathcal{P}_{T,M})$, showing that the system presents a natural time $T^*$ during which we should act upon it. This is supported by the fact that we need to assume that $T \geq T^*$ to have the existence of solutions and by the fact that in case $T > T^*$, there exist one or two time intervals of size $T - T^* > 0$ in which $u^*(t) = 0$. This motivates the introduction of a new problem, in which the final time $T$ is free. The functional we are interested in minimizing in this section is a convex combination of the cost $J$ used in the previous section and the final time:

$$\begin{cases} \inf\limits_{\substack{u \in \mathcal{U}_{T,M} \\ T>0}} J_\alpha(T, u), \\ p' = f(p) + ug(p), \quad p(0) = 0, \quad p(T) = \theta, \end{cases} \tag{$\mathcal{P}^\alpha_M$}$$

where $\mathcal{U}_{T,M}$ is given by (4.2), $\alpha \in [0, 1]$, and

$$J_\alpha(T, u) = (1 - \alpha) \int_0^T u(s)ds + \alpha T.$$

We expect to see the intervals where $u = 0$ disappear. Let us state the main result of this section.

**Theorem 4.3.1.** *Let $M > m^*$ and $\alpha > 0$. Then the unique solution of problem $(\mathcal{P}^\alpha_M)$ is $u^*(t) = M\mathbb{1}_{[0,T^*]}$ with $T^*$ given by (4.9). Thus the optimal value is $J(u^*) = T^*((1-\alpha)M+\alpha)$.*

*Proof.* The existence is proved separately in Appendix A. Here we focus on the characterization of the solution using first-order optimality conditions. We begin by excluding the case $\alpha = 1$, because the solution is trivial in this case. The problem reads

$$\begin{cases} \inf\limits_{\substack{u \in \mathcal{U}_{T,C,M} \\ T>0}} T, \\ p' = f(p) + ug(p), \quad p(0) = 0, \quad p(T) = \theta, \end{cases} \tag{$\mathcal{P}^1_{T,M}$}$$

and the solution is clearly the constant function $u^* = M$. Indeed, it is easy to prove that if the set $\{u^* < M\}$ has a positive measure, then we decrease the minimal time to

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

reach $\theta$ by increasing $u^*$. Moreover, by the reasoning carried out in Theorem 4.2.1 we know that $T^*$ is the time that it takes for the system

$$\begin{cases} p'_M = f(p_M) + Mg(p_M), \\ p_M(0) = 0, \end{cases}$$

to reach the point $p_M(T^*) = \theta$. The conclusion follows.

Next, let us assume that $\alpha < 1$, and let $(T, u^*)$ denote an optimal pair. To apply the Pontryagin maximum principle (PMP), we need to define the Hamiltonian of the system

$$\mathcal{H} : \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R} \times \{0, -1\} \times U \ni (t, p, q, q^0, u) = q(f(p) + ug(p)) + q^0(1 - \alpha)u,$$

where $q$ satisfies the equation

$$q' = -\frac{\partial \mathcal{H}}{\partial p}(t, p, q, q^0, u^*) = -q(f'(p) + u^* g'(p)),$$

and therefore $q$ has a constant sign.

The instantaneous maximization condition reads

$$u^*(t) \in \arg\max_{v \in U} \mathcal{H}(t, p, q, q^0, v) = \arg\max_{v \in U} (w(t) + q^0(1 - \alpha))v,$$

where $w(t) = q(t)g(p(t))$.

Thanks to a reasoning analogous to that in Theorem 4.2.1, we also get that $q$ is positive. Since the final time is free, we have the extra condition $u^*(T) = v_T$, where $v_T$ solves the one-dimensional optimization problem

$$\max_{v_T \in U} \mathcal{H}(T, \theta, q(T), q^0, v_T) = \max_{v_T \in U} (w(T) + q^0(1 - \alpha))v_T = -q^0\alpha.$$

This condition rules out the case $q^0 = 0$ because if we assumed that $q^0 = 0$, then we would have $\max_{0 \le v_T \le M} w(T)v_T = 0$. Since $w(t) > 0$, this would imply that we should have at the same time $v_T = M$ and $w(T)M = 0$, leading to a contradiction. Therefore $q^0 = -1$, and we infer that the first-order optimality conditions imply that

$$\begin{cases} w(t) \le 1 - \alpha \text{ on } \{u^* = 0\}, \\ w(t) = 1 - \alpha \text{ on } \{0 < u^* < M\}, \\ w(t) \ge 1 - \alpha \text{ on } \{u^* = M\}, \\ \max\limits_{0 \le v_T \le M} (w(T) - (1 - \alpha))v_T = \alpha. \end{cases} \tag{4.11}$$

Recall that $w$ increases if $p(t) \in (0, p^*)$ and decreases if $p(t) \in (p^*, \theta)$, where $p^*$ is given by (4.5). Mimicking the reasoning done in the proof of Theorem 4.2.1 (in the case

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

of fixed $T$), we can assert that $u^*$ is bang-bang and that the set $\{u^* = M\}$ is one single open interval. Since $\alpha > 0$, the last condition of (4.11) yields $u^*(T) > 0$, so $u^*(T) = M$. This, together with the fact that the time it takes the system to reach $p = \theta$ at speed $M$ is $T^*$, allows us to conclude that solutions must be of the form $u^*(t) = M\mathbb{1}_{[\xi,\xi+T^*]}$ with $\xi \geq 0$. Indeed, once the function has switched to $u^* = M$, it cannot switch back to $u^* = 0$. Then, looking at the functional we want to minimize, we conclude that $\xi = 0$, since the cost term is independent of $\xi$ and the term $\alpha T$ increases with respect to $\xi$. As a result, the (unique) solution is $u^* = M\mathbb{1}_{[0,T^*]}$. □

**Remark.** If we set $\alpha = 0$ in problem $(\mathcal{P}_M^\alpha)$, then we recover problem $(\mathcal{P}_{T,M})$ but without a restriction on the final time. This allows $T$ to go to infinity and explains that all the pairs of the form $(u_T, T)$ where

$$T > T^* \quad \text{and} \quad u_T(t) = M\mathbb{1}_{[\xi,\xi+T^*]}$$

with $\xi \in [0, T - T^*)$ solve this problem. Once the system has reached the final state $p(T) = \theta$, it can stay there indefinitely without using any mosquitoes (i. e., $u = 0$). It is not so realistic from a practical point of view and justifies fixing $T$ to avoid the emergence of such noncompact families of solutions.

## 4.4 Numerical simulations

This section is devoted to some numerical simulations. We will use the Python package GEKKO (see [5]), which solves, among other things, optimal control problems under constraints given by differential equations thanks to nonlinear programming solvers.

A particular attention has been paid to the development of a user-friendly source code, which is available online at

https://github.com/jesusbellver/Minimal-cost-time-strategies-for-mosquito-population-replacment

with the hope that it may serve as a useful basis for other research.

### 4.4.1 1D Case

Hereafter, we provide some simulations for the reduced problems $(\mathcal{P}_{T,M})$ and $(\mathcal{P}_M^\alpha)$, which can be seen as numerical confirmations of the theoretical results stated in Theorems 4.2.1 and 4.3.1. The parameters considered for these simulations are given in Table 4.1 using the biological parameters considered in [2].

Simulations for problem $(\mathcal{P}_{T,M})$ are shown in Figure 4.1. To deal with the constraint $p(T) = \theta$, we added a penalization term in the definition of the functional. As expected,

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Table 4.1:** Simulation parameter values considered in $(\mathcal{P}_{T,M})$ and $(\mathcal{P}_M^\alpha)$.

| Category | Parameter | Name | Value |
|---|---|---|---|
| Optimization | $T$ | Final time | 0.5 |
| | $M$ | Maximal instantaneous release rate | 10 |
| Biology | $b_1^0$ | Normalized wild birth rate | 1 |
| | $b_2^0$ | Normalized infected birth rate | 0.9 |
| | $d_1^0$ | Wild death rate | 0.27 |
| | $d_2^0$ | Infected death rate | 0.3 |
| | $K$ | Normalized carrying capacity | 1 |
| | $s_h$ | Cytoplasmatic incompatibility level | 0.9 |

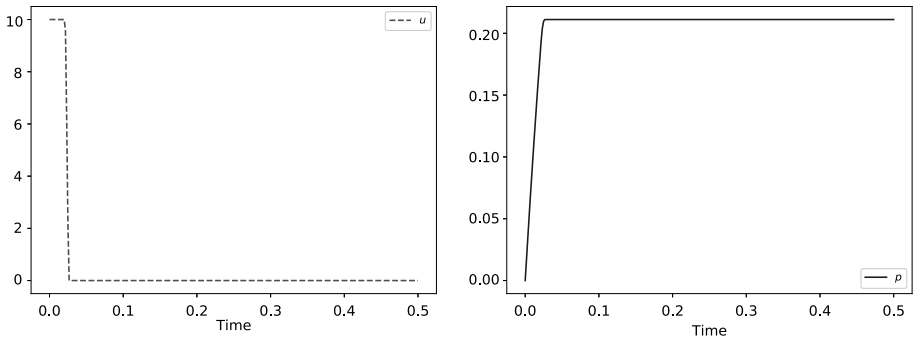



**Figure 4.1:** Simulation of the reduced problem $(\mathcal{P}_{T,M})$. The penalization parameter is $\varepsilon = 0.01$, and the number of elements in the ODE discretization is equal to 300.

we recover in Figure 4.1 that the optimal control is bang-bang and that all mosquitoes are only released for an interval of time. Although other optimal solutions may exist, we obtain a particular one, where the action is concentrated at the beginning of the time interval $[0, T]$.

Simulations for problem $(\mathcal{P}_M^\alpha)$ are provided in Figure 4.2. As proven in Theorem 4.3.1, the effect of letting the final time $T$ free and adding it with a weight to the functional we want to minimize leads to the absence of an interval during which no action is taken for every $\alpha \in (0, 1]$. In Figure 4.2, numerical solutions are plotted for $\alpha = 0.01$. The final time obtained is $T^*_{\text{num}} \approx 0.0238519$, which is very close to the expected theoretical one (which does not depend on $\alpha$)

$$T^* = \int_0^\theta \frac{dv}{f(v) + Mg(v)} \approx 0.0238122.$$

Results for other values of $\alpha$ are similar.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

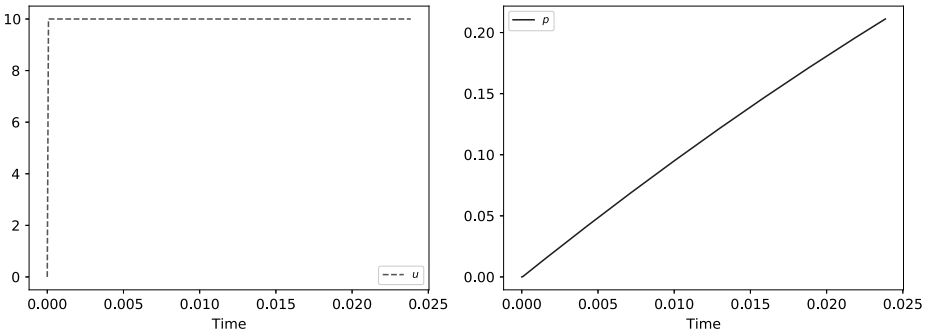



**Figure 4.2:** Simulation of the reduced problem ($\mathcal{P}_M^\alpha$) with $\alpha = 0.01$. The number of elements in the ODE discretization is equal to 300.

### 4.4.2 2D Case

In this section, we provide simulations for optimal control problems involving the full system (4.1). We will use the parameters of Table 4.2, where the biological parameters have been extracted from [19]. We choose $M$ to be ten times higher than the birth rate of wild mosquitoes in analogy with the simulations for the reduced problem.

**Table 4.2:** Values of simulation parameters for the optimal control problems involving the full system (4.1).

| Category | Parameter | Name | Value |
|---|---|---|---|
| Optimization | $M$ | Maximal instantaneous release rate | 112 |
| | $b_1$ | Wild birth rate | 11.2 |
| | $b_2$ | Infected birth rate | 10.1 |
| Biology | $d_1$ | Wild death rate | 0.04 |
| | $d_2$ | Infected death rate | 0.044 |
| | $K$ | Carrying capacity | 5124 |
| | $s_h$ | Cytoplasmatic incompatibility level | 0.9 |

To compute the carrying capacity $K$, we used the same procedure as in [19] but adapting it to our model. We will make our results relevant for an island of 74 ha with a mosquito density of 69 $\text{ha}^{-1}$, so the amount of wild mosquitoes at equilibrium is $n_1^* = 74 \times 69 = 5106$. Then, since $n_1^* = K(1 - \frac{d_1}{b_1})$, we obtain the following carrying capacity of the environment

$$K = \frac{n_1^*}{1 - \frac{d_1}{b_1}} \approx 5124.3011.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

We first deal with the case where $T$ is fixed; in other words, we solve the optimal control problem

$$\inf_{u\in\mathcal{U}_{T,M}} \int_0^T u(s)\,ds + \frac{1}{\varepsilon}\max\{n_1(T) - 10, n_2^* - 10 - n_2(T), 0\}, \tag{4.12}$$

where $\|\cdot\|_{\mathbb{R}^2}$ stands for the Euclidean norm in $\mathbb{R}^2$. The minimized criterion is a combination of the total amount of mosquitoes used and a penalization term standing for the final distance to the region $[0,10]\times[n_2^*-10, n_2^*]$, with $\varepsilon = 0.0001$.

It is easy to show that because of the pointwise constraint on the control function $u$, the steady state $(0, n_2^*)$ of system (4.1) cannot be reached in time $T$. This is why we chose to penalize the final distance to an arbitrary region that is clearly included in the basin of attraction of the steady state $(0, n_2^*)$ but is reachable.

The simulations are performed for different final times, and the results are given on Figure 4.3. We observe that if $T$ is not large enough to get sufficiently close to the

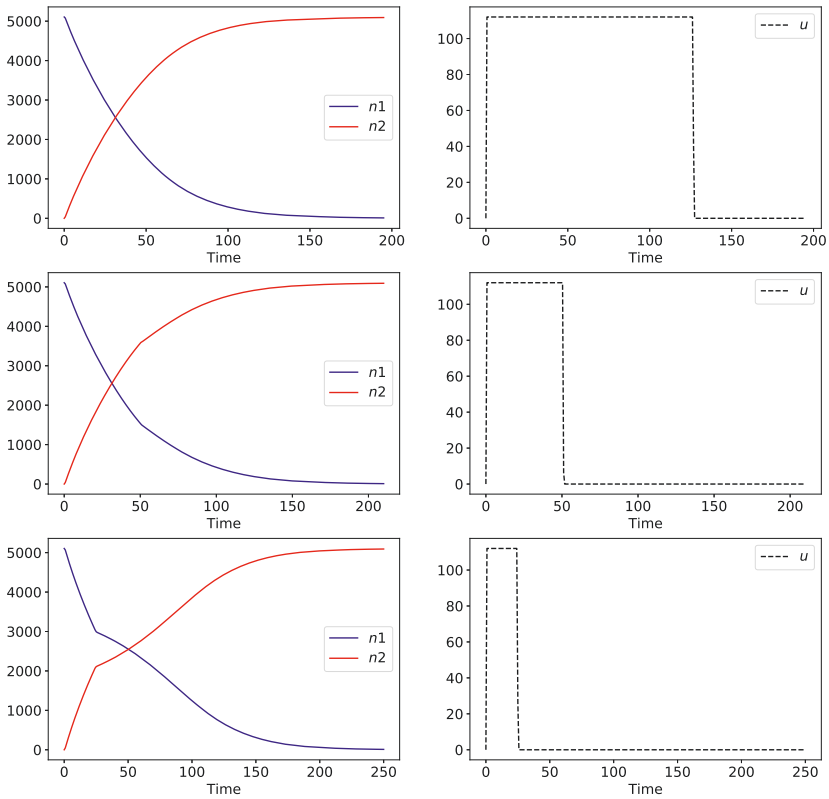



**Figure 4.3:** Simulation of the full problem (4.12) with $T$ fixed for $T = 195$ (first row), $T = 210$ (second row), and $T = 250$ (third row). The time step in the ODE discretization is $\Delta t = T/300$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

point $(0, n_2^*)$, then the control is the function $u$ equal to $M$ almost everywhere. As $T$ increases, the action is carried out in two stages: first, we have $u = M$ at least until the system enters the basin of attraction of the equilibrium point $(0, n_2^*)$; then $u = 0$ to let the system evolve without using mosquitoes. The larger $T$ is, the less it seems necessary to act. A possible explanation is that with only a little action, it is possible to enter the basin of attraction of $(0, n_2^*)$. Therefore, if $T$ is large, then we can stop acting early to decrease the amount of mosquitoes used. Otherwise, if $T$ is small, then we need to release a lot of mosquitoes because system would not get close enough to $(0, n_2^*)$ by itself.

Finally, on Figure 4.4, simulations are carried out for the full system (4.1) by letting $T$ free and replacing the previous cost functional by

$$u \mapsto (1-\alpha)\int_0^T u(t)dt + \alpha T + \frac{1}{\varepsilon}\max\{n_1(T) - 10, n_2^* - 10 - n_2(T), 0\}$$

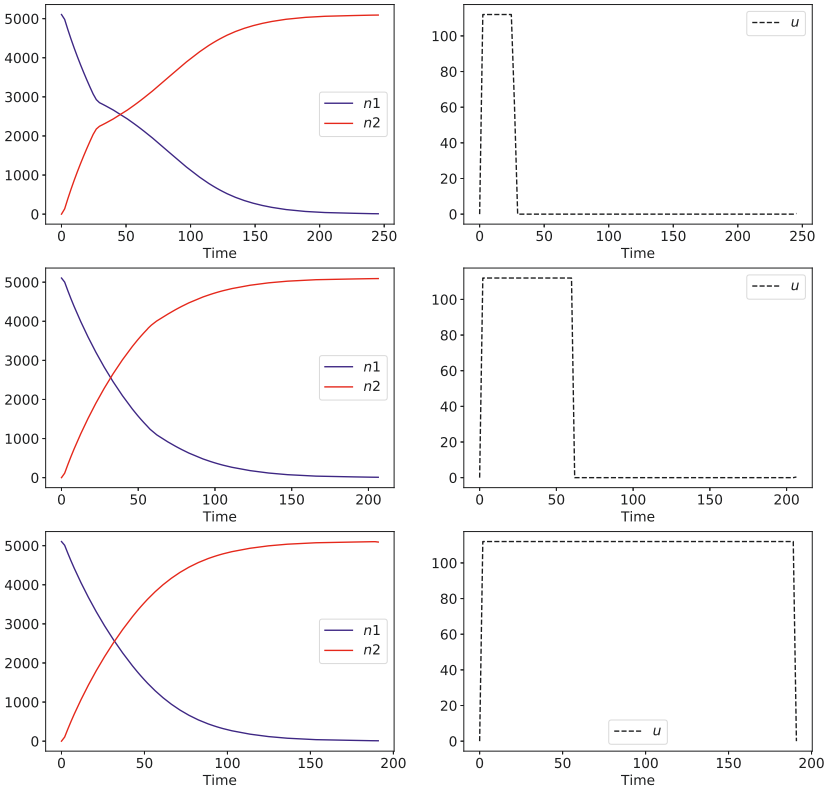



**Figure 4.4:** Simulation of the full problem with $T$ free for $\alpha = 0.1$ (first row), $\alpha = 0.5$ (second row), and $\alpha = 0.9$ (third row).The number of points in the ODE discretization is 101.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

with $\alpha \in [0,1]$. We see that the effect of increasing $\alpha$, and thus giving more importance to the time horizon $T$, has the same effect as decreasing $T$ in the case of fixed $T$. In these simulations the final times obtained are $T = 245.4$ for $\alpha = 0.1$, $T = 206.4$ for $\alpha = 0.5$, and $T = 190.9$ for $\alpha = 0.9$. The results obtained are very similar to those with fixed $T$ and very close to the duration during which the control is acting in that case.

# Appendix A. Existence of solutions for problem $(\mathcal{P}_{T,M})$

The set of admissible controls for problem $(\mathcal{P}_{T,M})$ is

$$\mathcal{D} = \{u \in \mathcal{U}_{T,M},\ p(T) \geq \theta\}.$$

Let us first prove that $\mathcal{D}$ is nonempty. This can be done by seeking under which assumptions we can ensure that constant controls $u(t) = u\mathbb{1}_{[0,T]}$ belong to the set $\mathcal{D}$. Let us introduce $p_{\bar{u}}$ solving

$$\begin{cases} p'_{\bar{u}} = f(p_{\bar{u}}) + \bar{u}g(p_{\bar{u}}) & \text{in } (0,T), \\ p_{\bar{u}}(0) = 0. \end{cases}$$

By integrating both sides of the differential equation we get that the time it takes for $p_{\bar{u}}$ to reach the point $\theta$, denoted $T_{\bar{u}}$, is

$$T_{\bar{u}} = \int_0^\theta \frac{dv}{f(v) + \bar{u}g(v)}.$$

Note that $T_{\bar{u}}$ is finite if we impose $\bar{u} > m^*$. Moreover, since we want $\bar{u} \in \mathcal{U}_{T,M}$, we have that $\bar{u} \leq M$, so $\bar{u} \in ]m^*, M]$. Finally, since the final time $T$ is fixed, we need $T_{\bar{u}} \leq T$, and using that $\int_0^\theta \frac{dv}{f(v)+\bar{u}g(v)}$ is decreasing with respect to $\bar{u}$, we deduce that $T^* \leq T_{\bar{u}}$. Thus we can conclude that $\mathcal{D}$ contains at least one constant control, and therefore it is nonempty if and only if $T^* \leq T$, as we assumed.

Since $\mathcal{D}$ is nonempty, we consider a minimizing sequence $(u_n)_{n\in\mathbb{N}} \in \mathcal{D}^{\mathbb{N}}$ for problem $(\mathcal{P}_{T,M})$. We have $0 \leq u_n \leq M$ a. e. for all $n \in \mathbb{N}$. Hence the sequence $(u_n)_{n\in\mathbb{N}}$ is uniformly bounded. Also, since $(L^1(0,T))' = L^\infty(0,T)$, using the Banach–Alaouglu theorem, we conclude that $\mathcal{D}$ is weakly-$\star$ compact, and therefore, up to a subsequence, $u_n \overset{*}{\underset{n\to\infty}{\rightharpoonup}} u^*$, i. e. $(u_n)_{n\in\mathbb{N}}$ converges in the weak-$\star$ topology of $L^\infty(0,T)$, and $0 \leq u^* \leq M$ a. e., so that $u^* \in \mathcal{D}$.

We now consider $(p_n)_{n\in\mathbb{N}}$ where $p_n$ solves $p'_n = f(p_n) + u_n g(p_n)$ with $p_n(0) = 0$. Since $f, g \in \mathcal{C}^\infty([0,1])$ and $0 \leq p_n \leq 1$, we deduce that $(p'_n)_{n\in\mathbb{N}}$ is bounded in $L^\infty(0,T)$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Hence $p_n \in \mathcal{C}^0([0,1])$, and therefore, using the Ascoli–Arzelà theorem, we conclude that, up to a subsequence, $p_n \xrightarrow{\mathcal{C}^0} p^*$ where $p^* \in W^{1,\infty}(0,T)$.

To conclude, since $u_n \underset{n\to\infty}{\rightharpoonup}^* u^*$, we have $\int_0^T \varphi u_n \to \int_0^T \varphi u^*$ for all $\varphi \in L^1(0,T)$. In particular, for $\varphi : t \mapsto 1$, we have $\int_0^T 1 \cdot u_n \to \int_0^T 1 \cdot u^*$. Hence $J(u^*) = \lim_{n\to\infty} J(u_n) = \inf_{u\in\mathcal{D}} J(u)$, and therefore problem $(\mathcal{P}_{T,M})$ admits a solution.

# Appendix B. Existence of solutions for problem $(\mathcal{P}_M^\alpha)$

To simplify the study of the existence of solutions and to avoid working on a variable domain, we make the following change of variables: $\tilde{p}(s) := p(Ts)$ and $\tilde{u}(s) := u(Ts)$, $s \in [0,1]$.

Then we are led to consider the problem

$$\begin{cases} \tilde{p}'(s) = T(f(\tilde{p}(s)) + \tilde{u}(s)g(\tilde{p}(s))), \quad \tilde{p}(0) = 0, \quad \tilde{p}(1) = \theta, \\ \inf\limits_{\tilde{u}\in L^\infty(0,1;[0,M])} \tilde{J}(T,\tilde{u}), \end{cases} \tag{$\tilde{\mathcal{P}}_M^\alpha$}$$

where $\tilde{J}(T,\tilde{u})$ is defined by

$$\tilde{J}(T,\tilde{u}) = (1-\alpha)\, T \int_0^1 \tilde{u}(s)ds + \alpha T.$$

Since the solutions of this system are the same as those of the system we are interested in, we will study the existence of solutions for the new one.

Let us define the set of $(T,\tilde{u})$ satisfying the constraints of problem $(\tilde{\mathcal{P}}_M^\alpha)$, i. e.,

$$\mathcal{D} := \{(T,\tilde{u}) \in \mathbb{R}^+ \times \mathcal{U}_{1,M} \times [0,1] \mid \tilde{p}(1) \geq \theta\}.$$

This set is clearly nonempty (consider, for instance, $T = T^*$ and $u(\cdot) = M\mathbb{1}_{[0,T^*]}(T^*\cdot)$). Consider a minimizing sequence $(T_n,\tilde{u}_n)_{n\in\mathbb{N}} \in \mathcal{D}^{\mathbb{N}}$ and let $\tilde{p}_n$ be the associated solution of the ODE defining problem $(\tilde{\mathcal{P}}_M^\alpha)$.

By minimality we have $\lim_{n\to\infty} \tilde{J}(T_n,\tilde{u}_n) < \infty$, i. e.,

$$\lim_{n\to\infty} (1-\alpha)\, T_n \int_0^1 \tilde{u}_n(s)ds + \alpha T_n < \infty.$$

Each term of the sum being bounded from below by 0, it follows that both of them are also bounded from above. Since $\alpha > 0$, $(T_n)_{n\in\mathbb{N}}$ is bounded, and therefore, up to a subsequence, $T_n \to \tilde{T} < \infty$. By mimicking the arguments used in Appendix A we get

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

that, up to a subsequence, $(\tilde{u}_n)_{n\in\mathbb{N}}$ converges to $\tilde{u}^* \in \mathcal{U}_{\tilde{T},M}$ weakly-$^*$ in $L^\infty(0,1;[0,M])$. Moreover, $(\tilde{p}_n)_{n\in\mathbb{N}}$ converges to $\tilde{p}^*$ in $\mathcal{C}^0([0,\tilde{T}])$, where $\tilde{p}^*$ solves the equation

$$(\tilde{p}^*)' = f(\tilde{p}^*) + \tilde{u}^* g(\tilde{p}^*) \quad \text{in } (0,\tilde{T}),$$

and $\tilde{p}^*(0) = 0$. As a consequence, $(\tilde{J}(T_n,\tilde{u}_n))_{n\in\mathbb{N}}$ converges to $\tilde{J}(\tilde{T},\tilde{u}^*)$, which concludes the proof.

# Bibliography


[1] L. Almeida, M. Duprez, Y. Privat, and N. Vauchelet. Mosquito population control strategies for fighting against arboviruses. *Math. Biosci. Eng.*, 16(6):6274–6297, 2019.

[2] L. Almeida, Y. Privat, M. Strugarek, and N. Vauchelet. Optimal releases for population replacement strategies: application to *Wolbachia*. *SIAM J. Math. Anal.*, 51(4):3170–3194, 2019.

[3] L. Alphey. Genetic control of mosquitoes. *Annu. Rev. Entomol.*, 59(1):205–224, 2014. PMID: 24160434.

[4] L. Alphey, M. Benedict, R. Bellini, G. G. Clark, D. A. Dame, M. W. Service, and S. L. Dobson. Sterile-insect methods for control of mosquito-borne diseases: an analysis. *Vector-Borne and Zoonotic Dis.*, 10(3):295–311, 2010.

[5] L. D. Beal, D. C. Hill, R. A. Martin, and J. D. Hedengren. Gekko optimization suite. *Processes*, 6(8):106, 2018.

[6] P.-A. Bliman, M. S. Aronna, F. C. Coelho, and M. A. H. B. da Silva. Ensuring successful introduction of *Wolbachia* in natural populations of *Aedes aegypti* by means of feedback control. *J. Math. Biol.*, 76(5):1269–1300, 2018.

[7] M. G. Guzman, S. B. Halstead, H. Artsob, P. Buchy, J. Farrar, D. J. Gubler, E. Hunsperger, A. Kroeger, H. S. Margolis, E. Martínez, et al. Dengue: a continuing global threat. *Nat. Rev. Microbiol.*, 8(12):S7–S16, 2010.

[8] A. Henrot and M. Pierre. *Shape Variation and Optimization*, volume 28 of Tracts in Mathematics. European Mathematical Society, Zürich, 2018.

[9] M. Hertig and S. B. Wolbach. Studies on Rickettsia-like micro-organisms in insects. *J. Med. Res.*, 44(3):329, 1924.

[10] A. A. Hoffmann, B. L. Montgomery, J. Popovici, I. Iturbe-Ormaetxe, P. H. Johnson, F. Muzzi, M. Greenfield, M. Durkan, Y. S. Leong, Y. Dong, H. Cook, J. Axford, A. G. Callahan, N. Kenny, C. Omodei, E. A. McGraw, P. A. Ryan, S. A. Ritchie, M. Turelli, and S. L. O'Neill. Successful establishment of *Wolbachia* in *Aedes* populations to suppress dengue transmission. *Nature*, 476(7361):454–457, Aug 2011.

[11] H. Hughes and N. F. Britton. Modelling the use of *Wolbachia* to control dengue fever transmission. *Bull. Math. Biol.*, 75(5):796–818, 2013.

[12] S. Kambhampati, K. S. Rai, and S. J. Burgun. Unidirectional cytoplasmic incompatibility in the mosquito, *Aedes albopictus*. *Evolution*, 47(2):673–677, 1993.

[13] L. A. Moreira, I. Iturbe-Ormaetxe, J. A. Jeffery, G. Lu, A. T. Pyke, L. M. Hedges, B. C. Rocha, S. Hall-Mendelin, A. Day, M. Riegler, et al.A *Wolbachia* symbiont in *Aedes aegypti* limits infection with dengue, chikungunya, and plasmodium. *Cell*, 139(7):1268–1278, 2009.

[14] L. Mousson, K. Zouache, C. Arias-Goeta, V. Raquin, P. Mavingui, and A.-B. Failloux. The native *Wolbachia* symbionts limit transmission of dengue virus in *Aedes albopictus*. *PLoS Negl. Trop. Dis.*, 6(12):e1989, 2012.


EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[15] M. Z. Ndii, R. I. Hickson, D. Allingham, and G. N. Mercer. Modelling the transmission dynamics of dengue in the presence of *Wolbachia*. *Math. Biosci.*, 262:157–166, 2015.
[16] G. Sallet and A. H. B. Silva Moacyr. Monotone dynamical systems and some models of *Wolbachia* in *Aedes aegypti* populations. *ARIMA Rev. Afr. Rech. Inform. Math. Appl.*, 20:145–176, 2015.
[17] J. G. Schraiber, A. N. Kaczmarczyk, R. Kwok, et al. Constraints on the use of lifespan-shortening *Wolbachia* to control dengue fever. *J. Theor. Biol.*, 297:26–32, 2012.
[18] S. P. Sinkins. Wolbachia and cytoplasmic incompatibility in mosquitoes. *Insect Biochem. Mol. Biol.*, 34(7):723–729, 2004.
[19] M. Strugarek, H. Bossin, and Y. Dumont. On the use of the sterile insect release technique to reduce or eliminate mosquito populations. *Appl. Math. Model.*, 68:443–470, 2019.
[20] A. P. Turley, L. A. Moreira, S. L. O'Neill, and E. A. McGraw. *Wolbachia* infection reduces blood-feeding success in the dengue fever mosquito, *Aedes aegypti*. *PLoS Negl. Trop. Dis.*, 3(9):e516, 2009.
[21] T. Walker, P. Johnson, L. Moreira, I. Iturbe-Ormaetxe, F. Frentiu, C. McMeniman, Y. S. Leong, Y. Dong, J. Axford, P. Kriesner, et al.The wMel *Wolbachia* strain blocks dengue and invades caged *Aedes aegypti* populations. *Nature*, 476(7361):450–453, 2011.
[22] WHO Department of Control Neglected Tropical Diseases, the WHO Department of Epidemic and Pandemic Alert and Response, and the Special Programme for Research and Training in Tropical Diseases. *Dengue: Guidelines for Diagnosis, Treatment, Prevention and Control*. World Health Organization, 2009.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Luis Almeida, Jorge Estrada, and Nicolas Vauchelet

# 5 The sterile insect technique used as a barrier control against reinfestation

**Abstract:** The sterile insect technique consists in massive release of sterilized males in the aim to reduce the size of mosquitoes population or even eradicate it. In this work, we investigate the feasibility of using the sterile insect technique as a barrier against reinvasion. More precisely, we provide some numerical simulations and mathematical results showing that performing the sterile insect technique on a band large enough may stop reinvasion.

**Keywords:** population dynamics, traveling waves, reaction–diffusion systems, supersolutions

**MSC 2010:** 35K57, 92D25, 35C07

## 5.1 Introduction

Due to the number of diseases that they transmit, mosquitoes are considered as one of the most dangerous animal species for humans. According to the World Health Organization [1], vector-borne diseases account for more than 17 % of all infectious diseases, causing more than 700 000 deaths annually. More than 3.9 billion people in over 128 countries are at risk of contracting dengue, with 96 millions cases estimated per year.

Several of the most prominent mosquito-transmitted diseases have no widely available vaccine; dengue, for instance, has only one licensed vaccine, Dengvaxia, with limited application [25]. Zika and chikungunya have no licensed vaccines, although several candidates are on diverse stages of development [24, 29]. Hence at this stage a natural strategy to control them is to act directly on the mosquito population. For this purpose, several strategies have been developed and experimented. Some techniques aim at replacing the existing population of mosquitoes by a population

**Acknowledgement:** The authors acknowledge the unknown referee for his/her comments and suggestions, which allowed us to improve this paper.

**Luis Almeida,** Sorbonne Université, CNRS, Université de Paris, Inria, Laboratoire Jacques-Louis Lions UMR7598, F-75005 Paris, France, e-mail: luis.almeida@sorbonne-universite.fr
**Jorge Estrada, Nicolas Vauchelet,** Laboratoire Analyse, Géométrie et Applications CNRS UMR 7539, Université Sorbonne Paris Nord, Villetaneuse, France, e-mails: estrada@math.univ-paris13.fr, vauchelet@math.univ-paris13.fr

https://doi.org/10.1515/9783110695984-005

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

unable to propagate the pathogens, the Wolbachia bacteria [15]. Other techniques aim at reducing the size of the mosquito population: the sterile insect technique (SIT) [4, 10], the release of insects carrying a dominant lethal (RIDL) [12, 14, 32], and the driving of antipathogen genes into natural populations [13, 19, 33]. Finally, other strategies combine both reduce and replace strategies [26].

In this paper, we focus on the SIT. This strategy has been introduced in the 1950s by Raymond C. Bushland and Edward F. Knipling. It consists of using area-wide releases of sterile insects to reduce reproduction in a field population of the same species. Indeed, wild female insects of the pest population do not reproduce when they are inseminated by released, radiation-sterilized males. For mosquito populations, this technique and the closely related incompatible insect technique (IIT) has been successfully used to drastically reduce their size in some isolated regions (see, e. g., [30, 34]). To predict the dynamics of mosquito populations, mathematical modeling is an important tool. In particular, there is a growing interest in the study of control strategies (see, e. g., [2, 5, 8, 18] and references therein). To conduct rigorous studies, these works usually neglect spatial dependency.

In fact, only few papers propose to incorporate the spatial variable in their study of SIT. In [17] the authors propose a simple scalar model to study the influence of the sterile insects density on the velocity of the spatial wave of spread of mosquitoes. The work [21] focuses on the influence of the release sites and the frequency of releases in the effectiveness of the sterile insect technique. In [27, 28] the authors conduct a numerical study on some mathematical models with spatial dependency to investigate the use of a barrier zone to prevent invasion of mosquitoes. However, up to our knowledge, there are no rigorous mathematical results on the existence of such a barrier zone. In this paper, we conduct a similar study as in [28] for another mathematical model, which was recently introduced in [31]. Moreover, we propose a strategy to rigorously prove the existence of a barrier zone under some conditions on the parameters.

The outline of the paper is the following: In Subsection 5.2.1, we introduce our dynamical system model and describe the variables and biological parameters, and also a simplified mode which results from additional assumptions. We analyze the existence of positive equilibria and the stability of the mosquito-free equilibrium. In Subsection 5.2.2, we introduce spatial models including diffusion. In Section 5.3, we perform numerical simulations for the spatial models to observe the existence of wave-blocking for a large enough release of sterile males. In Section 5.4, we prove the existence of a blocking phenomenon (some useful technical results are postponed to the Appendix). We end this paper with a conclusion.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

# 5.2 Mathematical model

## 5.2.1 Dynamical system

In the recent paper [31], the following mathematical model governing the dynamics of mosquitoes was proposed:

$$\begin{aligned}
\frac{dE}{dt} &= b\left(1-\frac{E}{K}\right)F-(\nu_E+\mu_E)E,\\
\frac{dM}{dt} &= (1-r)\nu_E E-\mu_M M,\\
\frac{dF}{dt} &= r\nu_E E\left(1-e^{-\beta(M+\gamma M_s)}\right)\frac{M}{M+\gamma M_s}-\mu_F F,\\
\frac{dM_s}{dt} &= u-\mu_s M_s.
\end{aligned} \tag{5.1}$$

In this system the mosquito population is divided into several compartments. The number of mosquitoes in the aquatic phase is denoted $E$; $M$ and $F$ denote, respectively, the numbers of adult wild males and adult females that have been fertilized; $M_s$ is the number of sterile mosquitoes that are released, the release function being denoted $u$. The fraction $\frac{M}{M+\gamma M_s}$ corresponds to the probability that a female mates with a wild male. Moreover, the term $(1-e^{-\beta(M+\gamma M_s)})$ has been introduced to model a strong Allee effect, that is, the fact that when the number of males is close to zero, it will be very difficult for a female find a wild male to mate with.

Finally, we have the following parameters:

- $b>0$ is the oviposition rate;
- $\mu_E>0$, $\mu_M>0$, $\mu_F>0$, and $\mu_s>0$ denote the death rates for the mosquitoes in the aquatic phase for wild adult males, adult females, and sterile males, respectively;
- $K$ is an environmental capacity for the aquatic phase, taking also into account the intraspecific competition;
- $\nu_E>0$ is the rate of emergence;
- $r\in(0,1)$ is the probability that a female emerges, and $(1-r)$ is the probability that a male emerges;
- $u$ is a control function corresponding to the number of sterile males that are released into the field.

Before introducing the spatial dependance in this system, we will first investigate the equilibria of (5.1) and their stability. For future use, we introduce the notation

$$\begin{aligned}
g(F,M_s) = {} & \frac{r\nu_E bF}{\frac{b}{K}F+\mu_E+\nu_E}\left(1-\exp\left(-\beta\left(\frac{(1-r)\nu_E bF}{\mu_M(\frac{b}{K}F+\mu_E+\nu_E)}+\gamma M_s\right)\right)\right)\\
& \times\frac{(1-r)\nu_E bF}{(1-r)\nu_E bF+\gamma\mu_M M_s(\frac{b}{K}F+\mu_E+\nu_E)}-\mu_F F.
\end{aligned} \tag{5.2}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Then we may use Lemma 3 in [31] to prove the following result.

**Proposition 5.2.1.** *Let* $N = \frac{br\nu_E}{\mu_F(\nu_E+\mu_E)}$ *and* $\psi = \frac{\mu_M}{(1-r)\nu_E K\beta}$. *Let us assume that* $N > \max\{1, 4\psi\}$. *Let* $\theta_0 \in (0,1)$ *be the unique solution of* $1-\theta_0 = -\frac{4\psi}{N}\ln\theta_0$ *and assume that*

$$\max_{\theta\in[\theta_0,1]}\left(-\ln\theta - \frac{1}{2\psi}\left(1-\sqrt{1+\frac{4\psi}{N}\frac{\ln(\theta)}{1-\theta}}\right)\right) > 0. \tag{5.3}$$

*Then we have:*

1. *When* $u = 0$, *there exist two positive equilibria for system* (5.1). *They are given by* $(\overline{E}_1, \overline{M}_1, \overline{F}_1, 0)$ *and* $(\overline{E}_2, \overline{M}_2, \overline{F}_2, 0)$ *with* $\overline{F}_1 < \overline{F}_2$ *and* $\overline{E}_i = \frac{b\overline{F}_i}{\frac{b}{K}\overline{F}_i+\nu_E+\mu_E}$, $\overline{M}_i = \frac{(1-r)\nu_E}{\mu_M}\overline{E}_i$ *for* $i = 1, 2$.
2. *There exists a positive constant* $\tilde{U}$ *large enough, such that if* $u = \overline{U}$ *is a constant such that* $\overline{U} > \tilde{U}$, *then the unique equilibrium for system* (5.1) *is the mosquito-free equilibrium* $(0, 0, 0, \overline{U}/\mu_s)$, *which is globally asymptotically stable.*

*Proof.* At equilibrium, we have

$$E = \frac{bF}{\frac{b}{K}F+\nu_E+\mu_E}, \quad M = \frac{(1-r)\nu_E}{\mu_M}E.$$

Injecting into the equilibrium equation for $F$, we deduce that any equilibrium should satisfy

$$g(F, M_s) = 0. \tag{5.4}$$

For $u = 0$, the only equilibrium of $\frac{dM_s}{dt} = u - \mu_s M_s$ is $M_s = 0$. Substituting $M_s = 0$ into (5.2), for (5.4), we obtain

$$g(F,0) = \frac{r\nu_E bF}{\frac{b}{K}F+\mu_E+\nu_E}\left(1-\exp\left(-\beta\left(\frac{(1-r)\nu_E bF}{\mu_M(\frac{b}{K}F+\mu_E+\nu_E)}\right)\right)\right) - \mu_F F = 0.$$

Note that $F = 0$ is always a solution. When $F > 0$, the equation $g(F,0) = 0$ may be rewritten

$$g_1(F) := r\nu_E b\left(1-\exp\left(-\beta\left(\frac{(1-r)\nu_E bF}{\mu_M(\frac{b}{K}F+\mu_E+\nu_E)}\right)\right)\right) = \mu_F\left(\frac{b}{K}F+\mu_E+\nu_E\right).$$

The right-hand side is an affine function. It is straightforward to verify that $g_1$ is an increasing concave function on $(0,+\infty)$ such that $g_1(0) = 0$ and $\lim_{F\to+\infty} g_1(F) = r\nu_E b(1-\exp(-\frac{\beta(1-r)\nu_E K}{\mu_M}))$. Then this concave function may intersect the affine function in 0, 1,

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

or 2 points. Note that a necessary (but not sufficient) condition for intersection is that $\lim_{F\to+\infty} g_1(F) > \mu_F(\mu_E + \nu_E)$. The fact that condition (5.3) guarantees the existence of two positive roots is proved in [31, Lemma 3].

For the second part, for a given constant $u = \overline{U}$, the only equilibrium of $\frac{dM_s}{dt} = u - \mu_s M_s$ is $M_s = \overline{U}/\mu_s$. Then injecting this relation into the expression for $g$ in (5.2), we deduce that for any $F \geq 0$, we have $\lim_{\overline{U}\to+\infty} g(F, \frac{\overline{U}}{\mu_s}) = -\mu_F F$. For any $\alpha \in (0, \mu_F)$, we may find $\overline{U}$ large enough such that for any $F \geq 0$, $g(F, \frac{\overline{U}}{\mu_s}) \leq -\alpha F$. Hence the only solution of the equation $g(F, \frac{\overline{U}}{\mu_s}) = 0$ is $F = 0$, which implies that the only equilibrium for system (5.1) is the mosquito-free equilibrium $(0, 0, 0, \overline{U}/\mu_s)$.

Moreover, the Jacobian at the mosquito-free equilibrium is given by the matrix

$$J(0,0,0,\overline{U}/\mu_s) = \begin{pmatrix} -(\nu_E + \mu_E) & 0 & b & 0 \\ (1-r)\nu_E & -\mu_M & 0 & 0 \\ 0 & 0 & -\mu_F & 0 \\ 0 & 0 & 0 & -\mu_s \end{pmatrix}.$$

We easily compute its eigenvalues given by $\{-\nu_E - \mu_E, -\mu_M, -\mu_F, -\mu_s\}$; they are all negative. Therefore the mosquito-free equilibrium is globally asymptotically stable. □

**Remark 5.2.2.** Although the proof does not give us an analytic formula for the equilibria, Remark 3.2 in [31] gives us a useful approximation of $\overline{M}_2$ when $\beta$ is not too small, as $\overline{M}_2 \approx \frac{1}{\psi\beta}(1 - \frac{1}{N})$.

## 5.2.2 Spatial model

To model the spatial dynamics, we suppose that adult mosquitoes diffuse according to a random walk. It is classical to model this active motion by adding a diffusion operator in the compartment of females and males ([6, 23]). We denote by $x$ the spatial variable. To simplify the approach, we only consider the one-dimensional case, $x \in \mathbb{R}$. Then all unknown functions depend on time $t > 0$ and position $x \in \mathbb{R}$. In this setting, model (5.1) becomes

$$\partial_t E = b\left(1 - \frac{E}{K}\right)F - (\nu_E + \mu_E)E, \tag{5.5a}$$

$$\partial_t M - D_u \partial_{xx} M = (1-r)\nu_E E - \mu_M M, \tag{5.5b}$$

$$\partial_t F - D_u \partial_{xx} F = r\nu_E E(1 - e^{-\beta(M+\gamma M_s)})\frac{M}{M + \gamma M_s} - \mu_F F, \tag{5.5c}$$

$$\partial_t M_s - D_u \partial_{xx} M_s = u - \mu_s M_s. \tag{5.5d}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

In this model, $D_u$ is a given diffusion coefficient, which is assumed to be the same for both male and female mosquitoes. This system is complemented with some nonnegative initial data denoted $E^0$, $M^0$, $F^0$, and $M_s^0 = 0$. Then there exists a nonnegative solution of (5.5). From now on, we will always assume that the parameters of the model are such that there are two positive equilibria for system (5.1) (see Proposition 5.2.1).

In the case where no control technique is implemented, i. e., $M_s = 0$, we expect that this model to predict the invasion of the whole domain by the wild mosquitoes. In order to illustrate this phenomenon, we perform some numerical simulations of system (5.5). This system is discretized thanks to a semiimplicit finite difference scheme on an uniform mesh. We use the numerical values in Table 5.1 taken from [31]. When $M_s = 0$, the numerical results are shown in Figure 5.1. We observe that there is a wave of mosquitoes that invades the whole domain. Hence, in the absence of control and in a homogeneous environment, it is expected that the solutions to (5.5) invade freely the spatial domain.

If we fix a constant value of the number of sterile males in the whole domain, then the invasion is expected to slow down. This observation has already been studied in [16]. Numerical illustrations of this phenomenon are provided in Figure 5.2. In this simulation the number of sterile males is fixed to be $M_s = 5000$ on the whole domain. This figure should be compared to Figure 5.1, where $M_s = 0$: we observe that the invasion is slowed down.

In this paper, we want to investigate the possibility of blocking the propagation of the spreading of mosquitoes by releasing sterile mosquitoes on a band of width $L$. Can the sterile insect technique be used to act as a barrier to avoid reinvasion of wild mosquitoes in a mosquito-free region? To answer this question, we first perform some numerical simulations in the next section.

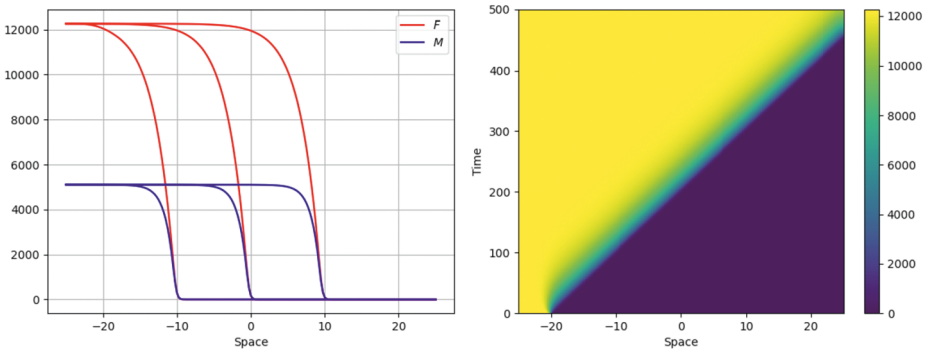



**Figure 5.1:** Numerical simulation of the time and space dynamics of $(F, M)$ solving (5.5) when $M_s = 0$. Left: profiles of solutions at three different times with same initial data; solutions are plotted at times $t = 100$, $t = 200$, and $t = 300$. Right: dynamics in time and space of the fertilized female density for system (5.5).

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

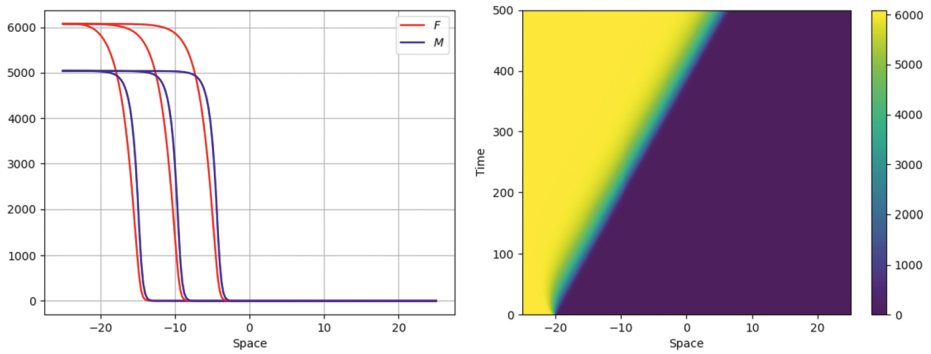



**Figure 5.2:** Numerical simulation of the time and space dynamics of $(F, M)$ solving (5.5) when $M_s = 5000$. Left: profiles of solutions at three different times with same initial data; solutions are plotted at times $t = 100$, $t = 200$, and $t = 300$. Right: dynamics in time and space of the females density for system (5.5). Comparing this figure with Figure 5.1, we observe that the invasion is slowing down.

## 5.3 Numerical simulations

We choose $u(t,x) = \overline{U}\mathbf{1}_{[0,L]}(x)$, where $\overline{U}$ is a given positive constant. We propose some numerical simulations. As before, we implement a finite difference scheme on an uniform mesh. The values of the numerical parameters are taken from [31] and are given in Table 5.1.

**Table 5.1:** Table of the numerical values used for the numerical simulations. These values are taken from [31].

| Parameter | $\beta$ | b | r | $\mu_E$ | $\nu_E$ | $\mu_F$ | $\mu_M$ | $\gamma$ | $\mu_s$ | $D_u$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Value | $10^{-2}$ | 10 | 0.49 | 0.03 | 0.05 | 0.04 | 0.1 | 1 | 0.12 | 0.0125 |
| Units | | $\text{day}^{-1}$ | | $\text{day}^{-1}$ | $\text{day}^{-1}$ | $\text{day}^{-1}$ | $\text{day}^{-1}$ | | $\text{day}^{-1}$ | $\text{km}^2 \cdot \text{day}^{-1}$ |

We show in Figure 5.3 the dynamics in time and space of the fertilized female population $F$ for model (5.5). The domain where the release of sterile males is performed is of width $L = 5\,\text{km}$. The release intensity is $\overline{U} = 10\,000\,\text{km}^{-1} \cdot \text{day}^{-1}$ (left), $\overline{U} = 20\,000\,\text{km}^{-1} \cdot \text{day}^{-1}$ (center), $\overline{U} = 30\,000\,\text{km}^{-1} \cdot \text{day}^{-1}$ (right). These results illustrate the influence of the intensity of the release $\overline{U}$ and of the width of the band on the efficiency of the blocking. We first notice that it seems that for $\overline{U}$ or $L$ large enough, the mosquito wave is not able to pass through the release zone. On the contrary, if $\overline{U}$ and $L$ are not large enough, the wave is only delayed by the release in the barrier zone.

It is also interesting to observe that as $\beta \to +\infty$, there is no blocking as it is illustrated in Figure 5.4 for model (5.5). To justify this observation, we observe that as $\beta \to +\infty$, the mosquito steady-state equilibrium is unstable for model (5.1). Indeed,

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

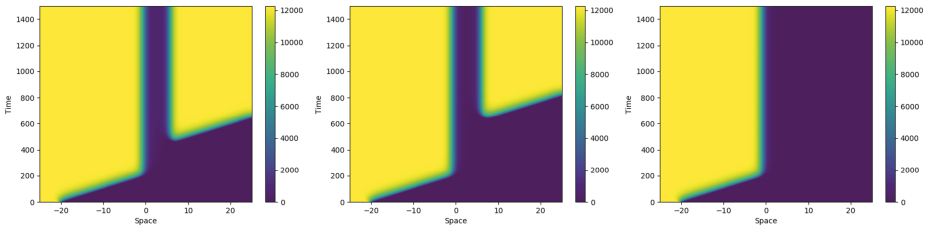



**Figure 5.3:** Numerical simulations of system (5.5) with $L = 5$ km and $\overline{U} = 10\,000\ \text{km}^{-1} \cdot \text{day}^{-1}$ (left), $\overline{U} = 20\,000\ \text{km}^{-1} \cdot \text{day}^{-1}$ (center), and $\overline{U} = 30\,000\ \text{km}^{-1} \cdot \text{day}^{-1}$ (right).

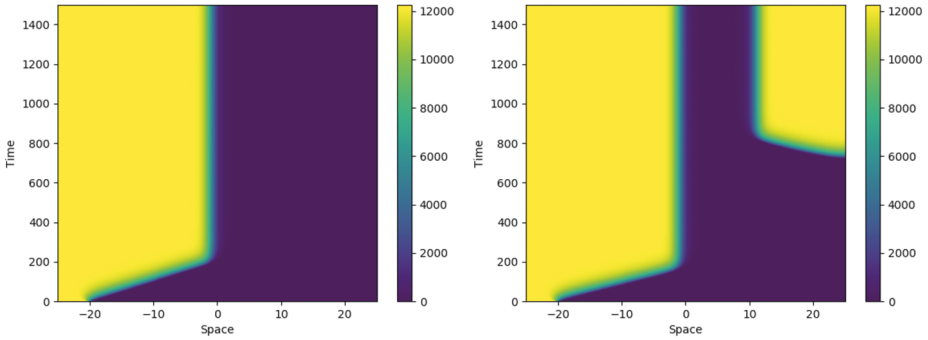



**Figure 5.4:** Numerical simulations for the model (5.5) with $\beta \to +\infty$, $L = 10$ km, and $\overline{U} = 40\,000\ \text{km}^{-1} \cdot \text{day}^{-1}$. In the figure on the left, $\beta = 10^{-2}$ as in Table 5.1, whereas on the right, we have the solution of the limit equation as $\beta \to +\infty$.

when $M_s = 0$ in system (5.1), after passing to the limit as $\beta \to +\infty$, the Jacobian matrix of the resulting system at the point $(0, 0, 0)$ is given by

$$\begin{pmatrix} -\nu_E - \mu_E & 0 & b \\ (1-r)\nu_E & -\mu_M & 0 \\ r\nu_E & 0 & -\mu_F \end{pmatrix}.$$

Clearly, $-\mu_M$ is one of the eigenvalues. For $br\nu_E > \mu_F(\nu_E + \mu_E)$, there is a positive real eigenvalue, and therefore the steady state would be unstable.

## 5.4 Mathematical approach

These numerical simulations seem to show that it is possible to block the spreading by releasing enough sterile males on a sufficiently wide domain. However, to be sure that it is not a numerical artifact and that the propagation is really blocked, we have to prove rigorous mathematical results. The study of blocking by local action has been done by several authors with applications, for instance, in biology or criminal studies

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[7, 9, 11, 17, 20, 22]. In this section, we will always assume that

$$u(t,x) = \overline{U}\mathbf{1}_{[0,L]}(x),$$

and we assume that (5.3) is satisfied, so that when $u = 0$, there are two stable steady states. We justify rigorously that for $L > 0$ and $\overline{U}$ large enough, we may block the propagation.

We first provide some useful bounds and comparisons on the variables of our system. Then we construct a stationary supersolution for (5.5) that vanishes at $+\infty$. When it exists, we will call this stationary supersolution a *barrier* for (5.5), since such a solution acts as a barrier to block the propagation.

## 5.4.1 Estimates

We first recall that since initial data are nonnegative, the solution $(E, F, M, M_s)$ of (5.5) is also nonnegative. Moreover, we have the following estimates.

**Lemma 5.4.1.** *Let us assume that $E^0 \le K$. Then, for any $t \ge 0$, we have $E \le K$. Moreover, $M$ and $F$ are uniformly bounded.*

*Proof.* The inequality is assumed to be true for $t = 0$. If there exist some $t_0 > 0$ and $x \in \mathbb{R}$ such that $E(t_0, x) \ge K$, then we deduce from the first equation in (5.5) that $\frac{d}{dt}E(t_0, x) < 0$, and hence $E(\cdot, x)$ is decreasing in the vicinity of $t_0$. Therefore, if this inequality is initially true, then it is true for any larger time. It then follows from a standard application of the maximum principle that $M$ and $F$ are also uniformly bounded. □

**Lemma 5.4.2.** *Let $u(t,x) = \overline{U}\mathbf{1}_{[0,L]}(x)$. Then the solution $M_s(t,x)$ to* (5.5d) *with initial data $M_s(t = 0) = 0$ converges uniformly on $x$ as $t \to +\infty$ to the solution $\overline{M}_s$ of the stationary equation*

$$-D_u\partial_{xx}\overline{M}_s = \overline{U}\mathbf{1}_{[0,L]} - \mu_s\overline{M}_s \quad \text{on } \mathbb{R}. \tag{5.6}$$

*Proof.* We have

$$\partial_t(M_s - \overline{M}_s) - \partial_{xx}(M_s - \overline{M}_s) = -\mu_s(M_s - \overline{M}_s), \quad M_s(t = 0) = 0.$$

Then, using the heat kernel, we deduce

$$M_s(t,x) = \overline{M}_s(x) - \int_{\mathbb{R}} \frac{1}{\sqrt{4\pi t}} e^{-\frac{y^2}{4t} - \mu_s t}\overline{M}_s(x - y)\, dy.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

The stationary solution $\overline{M}_s$ has the expression

$$\overline{M}_s(x) = \frac{\overline{U}}{2\sqrt{D_u \mu_s}} \int_0^L e^{-\sqrt{\frac{\mu_s}{D_u}}|x-y|}\, dy. \tag{5.7}$$

We can see from (5.7) that $\overline{M}_s$ is symmetric with respect to $x = \frac{L}{2}$ and decreasing on $(\frac{L}{2}, +\infty)$. Therefore $\overline{M}_s$ has a global maximum at $x = \frac{L}{2}$, which we compute to be $\overline{M}_s(\frac{L}{2}) = \frac{\overline{U}}{\mu_s}(1 - e^{-\sqrt{\frac{\mu_s}{D_u}}\frac{L}{2}})$, and we have

$$0 \le \overline{M}_s(x) - M_s(t,x) = \int_{\mathbb{R}} \frac{1}{\sqrt{4\pi t}} e^{-\frac{y^2}{4t} - \mu_s t} \overline{M}_s(x-y)\, dy \le e^{-\mu_s t} \overline{M}_s\left(\frac{L}{2}\right),$$

where we use the fact that the integral of the heat kernel equals 1. Therefore $M_s(t,x)$ converges to $\overline{M}_s$ exponentially in $t$ and uniformly in $x$. □

**Lemma 5.4.3.** *Let $L > 0$. Then, for any $L^* > 0$, the solution $\overline{M}_s$ of (5.6) satisfies, for any $x \in [0, L^*]$*

$$\overline{M}_s(x) \ge \frac{\overline{U}}{2\mu_s} \min\{1 - e^{-\sqrt{\frac{\mu_s}{D_u}}L}, e^{-\sqrt{\frac{\mu_s}{D_u}}L^*}(e^{\sqrt{\frac{\mu_s}{D_u}}L} - 1)\}.$$

*Proof.* Using the symmetry of $\overline{M}_s$, we have that for any $x \in [0, L^*]$, $\overline{M}_s(x) \ge \overline{M}_s(0) = \overline{M}_s(L)$ if $L > L^*$, and $\overline{M}_s(x) \ge \overline{M}_s(L^*)$ if $L \le L^*$. If $L > L^*$, then we see that

$$\overline{M}_s(0) = \frac{\overline{U}}{2\mu_s}(1 - e^{-\sqrt{\frac{\mu_s}{D_u}}L}),$$

and if $L \le L^*$, then we have that

$$\overline{M}_s(L^*) = \frac{\overline{U}}{2\mu_s} e^{-\sqrt{\frac{\mu_s}{D_u}}L^*}(e^{\sqrt{\frac{\mu_s}{D_u}}L} - 1).$$

□

An immediate consequence of this result is the following:

**Corollary 5.4.4.** *Let $\varepsilon > 0$ and $L^* > 0$. Then there exists $\overline{U}$ large enough such that, for any bounded nonnegative function $M$ and for any $x \in [0, L^*]$, we have*

$$(1 - e^{-\beta(M(x) + \gamma \overline{M}_s(x))}) \frac{M(x)}{M(x) + \gamma \overline{M}_s(x)} \le \varepsilon,$$

*where $\overline{M}_s$ is the solution of (5.6).*

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

## 5.4.2 Comparison principle

Let $\overline{M}_s \geq 0$ be a given nonnegative function. Let us consider the stationary problem associated with (5.5):

$$0 = b\left(1 - \frac{\overline{E}}{K}\right)\overline{F} - (\nu_E + \mu_E)\overline{E}, \tag{5.8a}$$

$$-D_u\overline{M}'' = (1-r)\nu_E\overline{E} - \mu_M\overline{M}, \tag{5.8b}$$

$$-D_u\overline{F}'' = r\nu_E\overline{E}(1 - e^{-\beta(\overline{M}+\gamma\overline{M}_s)})\frac{\overline{M}}{\overline{M}+\gamma\overline{M}_s} - \mu_F\overline{F}. \tag{5.8c}$$

We are particularly interested in stationary solutions that link the two stable equilibria $(0,0,0)$ and $(\overline{E}_2, \overline{M}_2, \overline{F}_2)$ defined in Proposition 5.2.1. Therefore we complement this system with conditions at infinity, $(\overline{E}, \overline{M}, \overline{F})(-\infty) = (\overline{E}_2, \overline{M}_2, \overline{F}_2)$ and $(\overline{E}, \overline{M}, \overline{F})(+\infty) = (0,0,0)$.

**Definition 5.4.5.** We call a *barrier* any supersolution of (5.8) such that $(\overline{E}, \overline{M}, \overline{F})(-\infty) = (\overline{E}_2, \overline{M}_2, \overline{F}_2)$ and $(\overline{E}, \overline{M}, \overline{F})(+\infty) = (0,0,0)$.

We recall that super-solutions are obtained by replacing the equality signs in (5.8) by the inequalities $\geq$.

The following result explains why we call such a solution a barrier: If the initial data is below this barrier, then the solution stays below the barrier, and hence the propagation (invasion of the whole domain by the wild population) will be blocked.

**Lemma 5.4.6.** *Let $M_s \geq \overline{M}_s$. Let us assume that the initial data are nonnegative and such that*

$$E^0 \leq K. \tag{5.9}$$

*Let us assume that there exists a barrier $(\overline{E}, \overline{M}, \overline{F})$ in the sense of Definition* 5.4.5.

*If we assume that $E^0 \leq \overline{E}$, $M^0 \leq \overline{M}$, and $F^0 \leq \overline{F}$, then the solution $(E, M, F)$ of* (5.5a)–(5.5c) *is such that for any $t \geq 0$ and $x \in \mathbb{R}$, we have $E(t,x) \leq \overline{E}$, $M(t,x) \leq \overline{M}$, and $F(t,x) \leq \overline{F}$.*

*Proof.* From Lemma 5.4.1 it follows that we may consider system (5.5a)–(5.5c) on the set $[0,K] \times \mathbb{R}^+ \times \mathbb{R}^+$. Let us denote

$$h(M, M_s) := \frac{M}{M + \gamma M_s}\left(1 - e^{-\beta(M+\gamma M_s)}\right).$$

We easily verify that on $\mathbb{R}^+ \times \mathbb{R}^+$, the map $h$ is nonincreasing with respect to $M_s$ and nondecreasing with respect to $M$. Therefore it is a monotone system, and we directly deduce the result by standard arguments.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

For completeness, we briefly explain these arguments. Let us denote $\delta E = E - \overline{E}$, $\delta M = M - \overline{M}$, and $\delta F = F - \overline{F}$. Subtracting and (5.8) from (5.5), we deduce

$$\begin{aligned}
\partial_t \delta E &\le b\Big(1 - \frac{E}{K}\Big)\delta F - \Big(\frac{b\overline{F}}{K} + \nu_E + \mu_E\Big)\delta E,\\
\partial_t \delta M - D_u \partial_{xx}\delta M &\le (1-r)\nu_E \delta E - \mu_M \delta M,\\
\partial_t \delta F - D_u \partial_{xx}\delta F &\le r\nu_E \delta E h(M, M_s) + r\nu_E \overline{E}\big(h(M, M_s) - h(\overline{M}, M_s)\big) - \mu_F \delta F,
\end{aligned}$$

where we use the fact that $h(\overline{M}, M_s) \le h(\overline{M}, \overline{M}_s)$. Denoting by $v_+ := \max\{v, 0\}$ the positive part of any real $v$, we multiply the first equation by $(\delta E)_+$, the second by $(\delta M)_+$, and the third by $(\delta F)_+$. Integrating over $\mathbb{R}$, we get

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}} (\delta E)_+^2\, dx &\le \int_{\mathbb{R}} b\Big(1 - \frac{E}{K}\Big)(\delta F)_+(\delta E)_+\, dx - \int_{\mathbb{R}} \Big(\frac{b\overline{F}}{K} + \nu_E + \mu_E\Big)(\delta E)_+^2\, dx,\\
\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}} (\delta M)_+^2\, dx + D_u \int_{\mathbb{R}} \big|\partial_x (\delta M)_+\big|^2\, dx &\le \int_{\mathbb{R}} (1-r)\nu_E (\delta E)_+(\delta M)_+\, dx - \int_{\mathbb{R}} \mu_M (\delta M)_+^2\, dx,\\
\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}} (\delta F)_+^2\, dx + D_u \int_{\mathbb{R}} \big|\partial_x (\delta F)_+\big|^2\, dx &\le \int_{\mathbb{R}} r\nu_E (\delta E)_+(\delta F)_+\, dx - \int_{\mathbb{R}} \mu_F (\delta F)_+^2\, dx\\
&\quad + \int_{\mathbb{R}} r\nu_E \overline{E}\big(h(M, M_s) - h(\overline{M}, M_s)\big)(\delta F)_+\, dx,
\end{aligned}$$

where we also use the fact that $h(M, M_s) \le 1$. Since $M \mapsto h(M, M_s)$ is increasing and uniformly Lipschitz-continuous, for some nonnegative constant $C$, we have

$$h(M, M_s) - h(\overline{M}, M_s) \le C(\delta M)_+.$$

Hence, adding these latter inequalities, we deduce that for some nonnegative constant $C$,

$$\frac{d}{dt}\int_{\mathbb{R}} \big((\delta E)_+^2 + (\delta F)_+^2 + (\delta M)_+^2\big)\, dx \le C \int_{\mathbb{R}} \big((\delta E)_+^2 + (\delta F)_+^2 + (\delta M)_+^2\big)\, dx.$$

We conclude by using the Grönwall lemma that from our assumption on the initial data we have that at time $t = 0$, $\delta E(t = 0) = 0$, $\delta M(t = 0) = 0$, and $\delta F(t = 0) = 0$. □

### 5.4.3 Existence of a barrier

Letting $\alpha \in (0, \mu_F)$, we deduce from Corollary 5.4.4 that for any $L > 0$, there exists $\overline{U}$ large enough such that for all $x \in [0, L]$, the solution $\overline{M}_s$ of (5.6) satisfies, for any non-

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

negative and bounded functions $\overline{M}$ and $\overline{F}$,

$$\frac{r\nu_E\overline{F}(x)}{\frac{b}{K}\overline{F}(x)+\mu_E+\nu_E}\left(1-e^{-\beta(\overline{M}(x)+\gamma\overline{M}_s(x))}\right)\frac{\overline{M}(x)}{\overline{M}(x)+\gamma\overline{M}_s(x)}\leq(\mu_F-\alpha)\overline{F}(x). \tag{5.10}$$

This suggests us to introduce the stationary problem

$$\overline{E}=\frac{b\overline{F}}{\frac{b}{K}\overline{F}+\mu_E+\nu_E}, \tag{5.11a}$$

$$-D_u\overline{M}''=(1-r)\nu_E\overline{E}-\mu_M\overline{M}, \tag{5.11b}$$

$$-D_u\overline{F}''=-\alpha\overline{F}\mathbf{1}_{[0,L]}+\left(r\nu_E\overline{E}(1-e^{-\beta\overline{M}})-\mu_F\overline{F}\right)\mathbf{1}_{[0,L]^c}. \tag{5.11c}$$

Note that this system is the stationary system for (5.5) when $M_s=0$ on $[0,L]^c$.

Any supersolution of (5.11) that satisfies the boundary conditions $(\overline{E},\overline{M},\overline{F})(-\infty)=(\overline{E}_2,\overline{M}_2,\overline{F}_2)$ and $(\overline{E},\overline{M},\overline{F})(+\infty)=(0,0,0)$ is a barrier, as defined in Definition 5.4.5. Then, to prove existence of a barrier, it suffices to prove that there exist supersolutions of (5.11), i. e., $(\overline{M},\overline{F})$ such that

$$-D_u\overline{M}''\geq\frac{(1-r)\nu_Eb\overline{F}}{\frac{b}{K}\overline{F}+\mu_E+\nu_E}-\mu_M\overline{M}\quad\text{on }\mathbb{R}\setminus\{0,L\}, \tag{5.12a}$$

$$-D_u\overline{F}''\geq-\alpha\overline{F}\mathbf{1}_{[0,L]}+\left(\frac{r\nu_Eb\overline{F}}{\frac{b}{K}\overline{F}+\mu_E+\nu_E}(1-e^{-\beta\overline{M}})-\mu_F\overline{F}\right)\mathbf{1}_{[0,L]^c}\quad\text{on }\mathbb{R}\setminus\{0,L\}, \tag{5.12b}$$

$$\overline{M}'(\xi^-)\geq\overline{M}'(\xi^+),\quad\overline{F}'(\xi^-)\geq\overline{F}'(\xi^+)\quad\text{for }\xi\in\{0,L\}, \tag{5.12c}$$

$$(\overline{M},\overline{F})(-\infty)=(\overline{M}_2,\overline{F}_2),\quad(\overline{M},\overline{F})(+\infty)=(0,0). \tag{5.12d}$$

We recall that $(\overline{M}_2,\overline{F}_2)$ is the larger mosquito equilibrium defined in (5.1). The main result of this work is the following:

**Theorem 5.4.7.** *Under the same assumptions as in Proposition* 5.2.1, *for any* $L>0$, *there exists* $\overline{U}$ *large enough such that there exists a solution* $(\overline{M},\overline{F})$ *of* (5.12).

*Hence there exists a barrier for system* (5.5).

*Proof.* Let $\alpha\in(0,\mu_F)$ and $L>0$ be fixed. We take $\overline{U}$ large enough such that (5.10) holds. The last point of the theorem is a consequence of the comparison principle. Therefore we just have to construct a supersolution of (5.11) that satisfies the boundary conditions. We construct this solution piecewise.

First, on $(-\infty,0)$, we recall that system (5.11) is a stationary system for (5.5) when $M_s=0$. From Proposition 5.2.1 we deduce that we may take $\overline{M}$, $\overline{F}$ on $(-\infty,0)$ as the larger values of the equilibria in 5.2.1, that is, $\overline{M}=\overline{M}_2$ and $\overline{F}=\overline{F}_2$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

On $(0,L)$, we introduce the notation $\tau := \frac{\nu_E b}{\mu_E+\nu_E}$ and take $(\overline{M}(L),\overline{F}(L))$ such that

$$0 < \overline{M}(L) < -\frac{1}{\beta}\ln\left(1-\frac{\mu_F}{2r\tau}\right), \quad 0 < \overline{F}(L) < \frac{\mu_M}{(1-r)\tau}\overline{M}(L). \tag{5.13}$$

Then we solve (5.11) on $(0,L)$ with boundary conditions $(\overline{M}(0),\overline{F}(0)) = (\overline{M}_2,\overline{F}_2)$ and $(\overline{M}(L),\overline{F}(L))$. From (5.11c) we deduce that

$$\overline{F}(x) = \frac{1}{2\sinh(\sqrt{\frac{\alpha}{D_u}}L)}((\overline{F}(L)-\overline{F}(0)e^{-\sqrt{\frac{\alpha}{D_u}}L})e^{\sqrt{\frac{\alpha}{D_u}}x} - (\overline{F}(L)-\overline{F}(0)e^{\sqrt{\frac{\alpha}{D_u}}L})e^{-\sqrt{\frac{\alpha}{D_u}}x}),$$

which is positive on $(0,L)$. From (5.11b) we deduce using the comparison principle that $\overline{M} \geq \underline{M} > 0$, where $\underline{M}$ is the solution of $-D_u\underline{M}'' + \mu_M\underline{M} = 0$ with $\underline{M}(0) = \overline{M}_2$ and $\underline{M}(L) = \overline{M}(L)$.

On $(L,+\infty)$, we first observe that for $F \geq 0$, we have $\frac{vbF}{\frac{b}{K}F+\nu_E+\mu_E} \leq \tau F$. Thus it suffices to prove that there exists a nonnegative solution on $(L,+\infty)$ tending to 0 at $+\infty$ of the system

$$\begin{aligned} -D_u\overline{M}'' &= (1-r)\tau\overline{F} - \mu_M\overline{M}, \\ -D_u\overline{F}'' &= (r\tau(1-e^{-\beta\overline{M}}) - \mu_F)\overline{F}, \end{aligned}$$

with $(\overline{M}(L),\overline{F}(L))$ as in (5.13). Notice that from (5.13) we have that (5.16) holds with $\phi(u) = r\tau(1-e^{-\beta u})$ (which is clearly increasing and Lipschitz continuous). Then we may apply the result in Proposition 5.A.1 in the Appendix: For $\overline{F}(L)$ small enough, there exists $(\overline{M}'(L^+),\overline{F}'(L^+))$ such that the solution of the above Cauchy problem is nonincreasing and nonnegative and goes exponentially to 0 as $x$ tends to $+\infty$. Thus our solution so far satisfies (5.12a) and (5.12b), and to conclude our proof, it only remains to prove (5.12c).

From the expression of $\overline{F}$ on $(0,L)$ we compute

$$\begin{aligned} \overline{F}'(0^+) &= \frac{\sqrt{\frac{\alpha}{D_u}}}{\sinh(\sqrt{\frac{\alpha}{D_u}}L)}\left(\overline{F}(L) - \overline{F}(0)\cosh\left(\sqrt{\frac{\alpha}{D_u}}L\right)\right), \\ \overline{F}'(L^-) &= \frac{\sqrt{\frac{\alpha}{D_u}}}{\sinh(\sqrt{\frac{\alpha}{D_u}}L)}\left(\overline{F}(L)\cosh\left(\sqrt{\frac{\alpha}{D_u}}L\right) - \overline{F}(0)\right). \end{aligned}$$

We already know that $\overline{F}'(0^-) = 0$, and from Lemma 5.A.2 we have $\overline{F}'(L^+) \leq -\sqrt{\frac{\mu_F}{2D_u}}\overline{F}(L)$. In the same manner, $\overline{M}'(0^-) = 0$, and we have from Lemma 5.A.3 that $\overline{M}'(L^+) \leq \frac{1}{\sqrt{\mu_M D_u}}((1-r)\tau\overline{F}(L) - \mu_M\overline{M}(L))$. By the comparison principle, $\overline{M}$ is bounded from above

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

by the solution on $(0, L)$ of

$$-D_u y'' + \mu_M y = \frac{(1-r)\nu_E b\overline{F}(0)}{\frac{b}{K}\overline{F}(0) + \nu_E + \mu_E} = \mu_M \overline{M}(0), \quad y(0) = \overline{M}(0), \quad y(L) = \overline{M}(L).$$

Then we have $\overline{M}'(0^+) \leq y'(0)$ and $\overline{M}'(L^-) \geq y'(L)$, where, after some straightforward computations,

$$y'(0) = \frac{\sqrt{\frac{\mu_M}{D_u}}}{\sinh(\sqrt{\frac{\mu_M}{D_u}}L)}(\overline{M}(L) - \overline{M}(0)), \quad y'(L) = \frac{\sqrt{\frac{\mu_M}{D_u}}\cosh(\sqrt{\frac{\mu_M}{D_u}}L)}{\sinh(\sqrt{\frac{\mu_M}{D_u}}L)}(\overline{M}(L) - \overline{M}(0)).$$

We can always find $L$ large enough and $(\overline{M}(L), \overline{F}(L))$, with $\overline{F}(L)$ small, such that (5.13) holds, $\overline{F}'(0^+) \leq 0$, $-\sqrt{\frac{\mu_F}{2D_u}}\overline{F}(L) \leq \overline{F}'(L^-)$, $y'(0) \leq 0$, and $y'(L) \geq \frac{1}{\sqrt{\mu_M D_u}}((1-r)\tau\overline{F}(L) - \mu_M\overline{M}(L))$. It concludes the construction of $(\overline{M}, \overline{F})$. □

Theorem 5.4.7 establishes that it is possible to block a propagating front of invading mosquitoes by performing a sterile insect technique on a domain wide enough with releases of a sufficiently large amount of sterile males. However, the mathematical proof presented above provides only a sufficient condition to block the front, and it is difficult to quantify the width of the domain and the number of sterile males. This is due to the fact that here we have to deal with a system of differential equations, whereas for a scalar equation, it is possible to obtain a better characterization of the barrier.

## 5.5 Conclusion

In this work, we have proved that mosquito invasion may be blocked by releasing a sufficient amount of sterile males in a band (see also [3] for a more detailed description of this result). The main motivation of this result is to use the sterile insect technique to build a sanitary cordon to protect a certain region (e. g., an urban area) from wild mosquitoes living in an exterior region (a reservoir area where they are abundant like, for instance, a forest) and which are vectors of a disease (like dengue, zika, or any other mosquito-borne disease). We remark that the mathematical result is not exclusive to mosquitoes. This technique of building a barrier to avoid invasion may be applied to other invading insect species (for instance, other disease vectors or agricultural pests). It is however important to characterize the population dynamics by an Allee effect such that the extinction steady state is stable.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

# Appendix. Existence of nonincreasing solution vanishing at infinity

In this appendix, we consider the second-order differential system on $(0, +\infty)$

$$-u'' = \alpha v - \beta u, \tag{5.14a}$$
$$-v'' = (\phi(u) - \mu)v, \tag{5.14b}$$
$$u(0) = u_0, \quad v(0) = v_0, \quad u'(0) = u_0', \quad v'(0) = v_0'. \tag{5.14c}$$

We assume that

$$\phi \text{ is nonnegative, nondecreasing, and Lipschitz continuous,} \tag{5.15}$$
$$\phi(u_0) \le \frac{\mu}{2}, \quad \alpha v_0 - \beta u_0 < 0. \tag{5.16}$$

Our main result is the following:

**Proposition 5.A.1.** *Under assumptions* (5.15) *and* (5.16), *for $v_0$ small enough, there exist $(u_0', v_0')$ such that the solution $(u, v)$ to* (5.14) *is such that $u$ and $v$ are nonincreasing and go exponentially to* 0 *as $x \to +\infty$.*

Before proving this result, we will prove two technical lemmas which concern (5.14a) and (5.14b) independently.

**Lemma 5.A.2.** *Let $\mu > 0$, let $\zeta$ be a nonincreasing and nonnegative function such that $\zeta(0) \le \frac{\mu}{2}$, and let $v_0 > 0$. Let us consider the Cauchy problem on* $(0, +\infty)$

$$-v'' = (\zeta(x) - \mu)v, \quad v(0) = v_0, \quad v'(0) = v_0'. \tag{5.17}$$

*Then there exists $v_0' \in [-\sqrt{\mu}v_0, -\sqrt{\frac{\mu}{2}}v_0]$ such that the solution of* (5.17) *is decreasing and tends to* 0 *at* $+\infty$. *Moreover, $v(x) \le Ce^{-\sqrt{\mu}x}$ for some nonnegative constant $C$.*

*Proof.* From the Duhamel formula we have that

$$\begin{aligned} v(x) &= v_0 \cosh(\sqrt{\mu}x) + \frac{v_0'}{\sqrt{\mu}} \sinh(\sqrt{\mu}x) - \frac{1}{\sqrt{\mu}} \int_0^x \zeta(z)v(z) \sinh(\sqrt{\mu}(x - z))\, dz \\ &\le v_0 \cosh(\sqrt{\mu}x) + \frac{v_0'}{\sqrt{\mu}} \sinh(\sqrt{\mu}x) \end{aligned} \tag{5.18}$$

as long as $v \ge 0$. Hence, we deduce that if $v_0' < -\sqrt{\mu}v_0$, then the right-hand side vanishes, which implies that $v$ vanishes. From the assumptions on $\zeta$ we also have

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

that

$$\begin{aligned} v(x) &= v_0 \cosh\Big(\sqrt{\frac{\mu}{2}}x\Big) + v_0' \sqrt{\frac{2}{\mu}} \sinh\Big(\sqrt{\frac{\mu}{2}}x\Big) \\ &\quad - \sqrt{\frac{2}{\mu}} \int_0^x \Big(\zeta(z) - \frac{\mu}{2}\Big) v(z) \sinh\Big(\sqrt{\frac{\mu}{2}}(x-z)\Big)\, dz \\ &\geq v_0 \cosh\Big(\sqrt{\frac{\mu}{2}}x\Big) + v_0' \sqrt{\frac{2}{\mu}} \sinh\Big(\sqrt{\frac{\mu}{2}}x\Big). \end{aligned}$$

If $v_0' > -\sqrt{\frac{\mu}{2}} v_0$, then we have $v > 0$, and $v(x)$ goes to $+\infty$ as $x \to +\infty$.

Since we look for a decreasing solution $v$, we may invert it into a function $x(v)$. Let us denote $w(v) = -v'(x(v))$. Then we have

$$w'(v) = \frac{(\mu - \zeta(x(v)))v}{w(v)}, \quad w(v_0) = -v_0'.$$

The question is to know whether there exists $v_0' < 0$ such that the solution of this equation is defined on $(0, v_0)$ and such that $w(0) = 0$.

There are two possibilities: either $w$ vanishes on $v_c \in (0, v_0)$, and then the solution is defined only on $(v_c, v_0)$, and we say that this solution is type I, or $w$ does not vanish on $(0, v_0)$, and then we set $v_c = 0$, and we say that this solution is type II. We look for a solution that is both type I and type II.

Clearly, if $v_0' < \overline{v}_0'$, then by the Cauchy–Lipschitz theorem the corresponding solutions satisfy $w < \overline{w}$. Therefore the map $v_0' \mapsto v_c$ is nonincreasing and continuous. Moreover, we saw at the beginning of this proof that if $v_0' > -\sqrt{\frac{\mu}{2}} v_0$, then the solution is type I, and if $v_0' < -\sqrt{\mu} v_0$, then the solution is type II. Thus by the continuity there exists $v_0'$ such that $v_c = 0$ and $w(0) = 0$. Moreover, we also have $v_0' \in [-\sqrt{\mu} v_0, -\sqrt{\frac{\mu}{2}} v_0]$, and with (5.18), we deduce the estimate $v(x) \leq \frac{1}{2}(v_0 + \frac{v_0'}{\sqrt{\mu}}) e^{-\sqrt{\mu}x}$. □

In the same spirit, we have the following result.

**Lemma 5.A.3.** *Let $v$ be a nonincreasing nonnegative function on $[0, \infty)$. Let $u_0 > 0$. We assume that $\alpha v(0) < \beta u_0$. We consider the Cauchy problem on $(0, +\infty)$*

$$-u'' = \alpha v - \beta u, \quad u(0) = u_0, \quad u'(0) = u_0'. \tag{5.19}$$

*Then there exists $u_0' \in [-\sqrt{\beta} u_0, \frac{1}{\sqrt{\beta}}(\alpha v(0) - \beta u_0)]$ such that the solution of (5.19) is decreasing and tends to 0 at $+\infty$. Moreover, $u(x) \leq Ce^{-\sqrt{\beta}x}$ for some nonnegative constant $C$.*

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*Proof.* Following the idea of the proof of Lemma 5.A.2, it remains to prove that there exist type I and type II solutions. From the Duhamel formula we have

$$u(x) = u_0 \cosh(\sqrt{\beta}x) + \frac{u_0'}{\sqrt{\beta}} \sinh(\sqrt{\beta}x) - \frac{\alpha}{\sqrt{\beta}} \int_0^x v(z) \sinh(\sqrt{\beta}(x-z))\, dz$$
$$\leq u_0 \cosh(\sqrt{\beta}x) + \frac{u_0'}{\sqrt{\beta}} \sinh(\sqrt{\beta}x),$$

where we used the fact that $v$ is nonnegative. Hence the solution is type I if $u_0' < -\sqrt{\beta}u_0$. Moreover, using the fact that $v(x) \leq v(0)$, we have

$$u(x) \geq u_0 \cosh(\sqrt{\beta}x) + \frac{u_0'}{\sqrt{\beta}} \sinh(\sqrt{\beta}x) + \frac{\alpha}{\beta} v(0)(1 - \cosh(\sqrt{\beta}x))$$
$$= \frac{\alpha}{\beta} v(0) + \left(u_0 - \frac{\alpha}{\beta} v(0) + \frac{u_0'}{\sqrt{\beta}}\right) e^{\sqrt{\beta}x} + \left(u_0 - \frac{\alpha}{\beta} v(0) - \frac{u_0'}{\sqrt{\beta}}\right) e^{\sqrt{\beta}x}.$$

Therefore the solution is type II if $u_0' > \frac{1}{\sqrt{\beta}}(\alpha v(0) - \beta u_0)$.

By a continuity and monotonicity argument, as in the proof of Lemma 5.A.2, there exists $u_0' \in [-\sqrt{\beta}u_0, \frac{1}{\sqrt{\beta}}(\alpha v(0) - \beta u_0)]$ such that the solution of (5.19) is such that $\lim_{x\to+\infty} u(x) = 0$. Moreover, we have from the first estimate in this proof that $u(x) \leq \frac{1}{2}(u_0 + \frac{u_0'}{\sqrt{\beta}}) e^{\sqrt{\beta}x}$. □

*Proof of Proposition* 5.A.1. Let us consider a nonincreasing nonnegative function $u$ on $[0, +\infty)$ such that $u(0) = u_0$. Then by Lemma 5.A.2 there exists $v_0'$ such that the solution of

$$-v'' = (\phi(u) - \mu)v, \quad v(0) = v_0, \quad v'(0) = v_0',$$

is nonincreasing and nonnegative and decays exponentially to 0 at $+\infty$. With such a function $v$, by Lemma 5.A.3 there exists $u_0'$ such that the solution of (5.19) is nonincreasing and nonnegative and decays exponentially to 0 at $+\infty$. We denote by $\mathcal{F}(u)$ this solution, which allows us to define the map $\mathcal{F}$. Let us denote by $\mathcal{A}$ the closed subset of functions in $H^1(0, +\infty)$ that are nonnegative and nonincreasing and take the value $u_0$ at 0. Clearly, $\mathcal{F}$ maps $\mathcal{A}$ into itself. Let us prove that $\mathcal{F}$ is a contraction on $\mathcal{A}$ for $v_0$ small enough.

Let $u_1$ and $u_2$ be in $\mathcal{A}$. Then by the definition of $\mathcal{F}$ we have

$$-(\mathcal{F}(u_1) - \mathcal{F}(u_2))'' + \beta(\mathcal{F}(u_1) - \mathcal{F}(u_2)) = \alpha(v_1 - v_2), \tag{5.20}$$
$$-(v_1 - v_2)'' = (\phi(u_1) - \mu)(v_1 - v_2) + v_2(\phi(u_1) - \phi(u_2)). \tag{5.21}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

On the one hand, multiplying (5.20) by $(\mathcal{F}(u_1) - \mathcal{F}(u_2))$ and integrating, after an integration by parts, we obtain

$$\int_0^\infty \left|\left(\mathcal{F}(u_1) - \mathcal{F}(u_2)\right)'\right|^2 dx + \beta \int_0^\infty \left(\mathcal{F}(u_1) - \mathcal{F}(u_2)\right)^2 dx = \int_0^\infty \alpha(v_1 - v_2)\left(\mathcal{F}(u_1) - \mathcal{F}(u_2)\right) dx.$$

By the Cauchy–Schwarz inequality we deduce

$$\left\|\mathcal{F}(u_1) - \mathcal{F}(u_2)\right\|_{H^1(0,+\infty)} \leq \frac{\alpha}{\min(1,\beta)} \|v_1 - v_2\|_{L^2(0,+\infty)}. \tag{5.22}$$

On the other hand, multiplying (5.21) by $(v_1 - v_2)$ and integrating, we get, in a similar way,

$$\begin{aligned}\int_0^\infty \left|(v_1 - v_2)'\right|^2 dx &= \int_0^\infty (\phi(u_1) - \mu)|v_1 - v_2|^2 dx + \int_0^\infty v_2(\phi(u_1) - \phi(u_2))(v_1 - v_2)\, dx \\ &\leq -\frac{\mu}{2} \int_0^\infty |v_1 - v_2|^2 dx + v_0 \|\phi'\|_\infty \int_0^\infty |u_1 - u_2||v_1 - v_2|\, dx,\end{aligned}$$

where we use (5.16), the Lipschitz continuity of $\phi$ (see (5.15)), and the fact that $v_2$ is decreasing. Applying again the Cauchy–Schwarz inequality, we obtain

$$\|v_1 - v_2\|_{L^2(0,+\infty)} \leq \frac{2v_0\|\phi'\|_\infty}{\mu} \|u_1 - u_2\|_{L^2(0,+\infty)}.$$

With (5.22), we conclude that

$$\left\|\mathcal{F}(u_1) - \mathcal{F}(u_2)\right\|_{H^1(0,+\infty)} \leq \frac{2v_0\alpha\|\phi'\|_\infty}{\mu \min(1,\beta)} \|v_1 - v_2\|_{L^2(0,+\infty)}.$$

Hence $\mathcal{F}$ is a contraction for $v_0$ small enough. Thus there exists a unique fixed point, which is a solution of our problem. □

# Bibliography

[1] https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases.

[2] L. Almeida, M. Duprez, Y. Privat, and N. Vauchelet. Control strategies on mosquitos population for the fight against arboviruses. *Math. Biosci. Eng.*, 16(6):6274–6297, 2019.

[3] L. Almeida, J. Estrada, and N. Vauchelet. Wave blocking in a mosquito population model with introduced sterile males. Pre-print https://hal.archives-ouvertes.fr/hal-03377080.

[4] L. Alphey, M. Q. Benedict, R. Bellini, G. G. Clark, D. Dame, M. Service, and S. Dobson. Sterile-insect methods for control of mosquito-borne diseases: an analysis. *Vector Borne Zoonotic Dis.*, 10:295–311, 2010.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[5] R. Anguelov, Y. Dumont, and J. Lubuma. Mathematical modeling of sterile insect technology for control of *anopheles* mosquito. *Comput. Math. Appl.*, 64:374–389, 2012.
[6] D. G. Aronson and H. F. Weinberger. Nonlinear diffusion in population genetics, combustion, and nerve pulse propagation. In J. A. Goldstein, editor, *Partial Differential Equations and Related Topics*, volume 446 of Lecture Notes in Mathematics. Springer, Berlin, Heidelberg, 1975.
[7] H. Berestycki, N. Rodríguez, and L. Ryzhik. Traveling wave solutions in a reaction–diffusion model for criminal activity. *Multiscale Model. Simul.*, 11(4):1097–1126, 2013.
[8] P.-A. Bliman, D. Cardona-Salgado, Y. Dumont, and O. Vasilieva. Implementation of control strategies for sterile insect techniques. *Math. Biosci.*, 314:43–60, 2019.
[9] G. Chapuisat and R. Joly. Asymptotic profiles for a traveling front solution of a biological equation, *Math. Models Methods Appl. Sci.*, 21(10):2155–2177 2011.
[10] V. A. Dyck, J. Hendrichs, and A. S. Robinson. *Sterile Insect Technique Principles and Practice in Area-Wide Integrated Pest Management*. Springer, 2005.
[11] S. Eberle. Front blocking versus propagation in the presence of a drift term in the direction of propagation, *Nonlinear Anal.*, 197:111836, 2020.
[12] G. Fu, et al. Female-specific flightless phenotype for mosquito control. *Proc. Natl. Acad. Sci.*, 107(10):4550–4554, 2010.
[13] F. Gould, Y. Huang, M. Legros, and A. L. Lloyd. A Killer Rescue system for self-limiting gene drive of anti-pathogen constructs. *Proc. R. Soc. B*, 275:2823–2829, 2008.
[14] J. Heinrich and M. Scott. A repressible female-specific lethal genetic system for making trans-genic insect strains suitable for a sterile-release program. *Proc. Natl. Acad. Sci. USA*, 97(15):8229–8232, 2000.
[15] A. A. Hoffmann, et al. Successful establishment of *Wolbachia* in *Aedes* populations to suppress dengue transmission. *Nature*, 476(7361):454–457, 2011.
[16] M. A. Lewis and P. van den Driessche. Waves of extinction from sterile insect release. *Math. Biosci.*, 116(2):221–247, 1993.
[17] T. J. Lewis and J. P. Keener. Wave-block in excitable media due to regions of depressed excitability. *SIAM J. Appl. Math.*, 61:293–316, 2000.
[18] J. Li and Z. Yuan. Modelling releases of sterile mosquitoes with different strategies. *J. Biol. Dyn.*, 9:1–14, 2015.
[19] J. M. Marshall, G. W. Pittman, A. B. Buchman, and B. A. Hay. Semele: a killer-male, rescue-female system for suppression and replacement of insect disease vector populations. *Genetics*, 187(2):535–551, 2011.
[20] G. Nadin, M. Strugarek, and N. Vauchelet. Hindrances to bistable front propagation: application to *Wolbachia* invasion. *J. Math. Biol.*, 76(6):1489–1533, 2018.
[21] T. P. Oléron Evans and S. R. Bishop. A spatial model with pulsed releases to compare strategies for the sterile insect technique applied to the mosquito *Aedes aegypti*. *Math. Biosci.*, 254:6–27, 2014.
[22] J. Pauwelussen. One way traffic of pulses in a neuron. *J. Math. Biol.*, 15:151–171, 1982.
[23] B. Perthame. *Parabolic Equations in Biology*. Lecture Notes on Mathematical Modelling in the Life Sciences. Springer International Publishing, 2015.
[24] G. A. Poland, I. G. Ovsyannikova, and R. B. Kennedy. Zika vaccine development: current status. *Mayo Clin. Proc.*, 94(12):2572–2586, 2019.
[25] M. Redoni, S. Yacoub, L. Rivino, et al. Dengue: status of current and under-development vaccines. *Rev. Med. Virol.*, 30:e2101, 2020. https://doi.org/10.1002/rmv.2101.
[26] M. A. Robert, K. Okamoto, F. Gould, and A. L. Lloyd. A reduce and replace strategy for suppressing vector-borne diseases: insights from a deterministic model. *PLoS ONE*, 8:e73233, 2012.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[27] S. Seirin Lee, R. E. Baker, E. A. Gaffney, and S. M. White. Modelling *Aedes aegypti* mosquito control via transgenic and sterile insect techniques: endemics and emerging outbreaks. *J. Theor. Biol.*, 331:78–90, 2013.
[28] S. Seirin Lee, R. E. Baker, E. A. Gaffney, and S. M. White. Optimal barrier zones for stopping the invasion of *Aedes aegypti* mosquitoes via transgenic or sterile insect techniques. *Theor. Ecol.*, 6:427–442, 2013.
[29] Gao Shan, Song Siqi, and Zhang Leiliang. Recent progress in vaccine development against Chikungunya virus. *Front. Microbiol.*, 2019. https://doi.org/10.3389/fmicb.2019.02881.
[30] B. Stoll, H. Bossin, H. Petit, J. Marie, and M. Cheong Sang. Suppression of an isolated population of the mosquito vector *Aedes polynesiensis* on the atoll of Tetiaroa, French Polynesia, by sustained release of *Wolbachia*-incompatible male mosquitoes. In *Conference: ICE – XXV International Congress of Entomology*, At Orlando, Florida, USA.
[31] M. Strugarek, H. Bossin, and Y. Dumont. On the use of the sterile insect technique or the incompatible insect technique to reduce or eliminate mosquito populations. *Appl. Math. Model.*, 68:443–470, 2019.
[32] D. D. Thomas, C. A. Donnelly, R. J. Wood, and L. S. Alphey. Insect population control using a dominant, repressible, lethal genetic system. *Science*, 287(5462):2474–2476, 2000.
[33] C. M. Ward, J. T. Su, Y. Huang, A. L. Lloyd, F. Gould, and B. A. Hay. Medea selfish genetic elements as tools for altering traits of wild populations: a theoretical analysis. *Evolution*, 65(4):1149–1162, 2011.
[34] X. Zheng, et al. Incompatible and sterile insect techniques combined eliminate mosquitoes. *Nature*, 572(7767):56–61, 2019.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Evelyn Herberg and Michael Hinze

# 6 Variational discretization approach applied to an optimal control problem with bounded measure controls

**Abstract:** We consider a parabolic optimal control problem with initial measure control. The cost functional consists of a tracking term corresponding to the observation of the state at final time. Instead of a regularization term in the cost functional, we follow [6] and consider a bound on the measure norm of the initial control. The variational discretization of the problem, together with the optimality conditions, induces maximal discrete sparsity of the initial control, i. e., Dirac measures in space. We present numerical experiments to illustrate our approach.

**Keywords:** variational discretization, optimal control, sparsity, partial differential equations, measures

**MSC 2010:** 49M25, 65K10

## 6.1 Introduction

We consider the following optimal control problem analyzed in [6]:

$$\min_{u\in U_\alpha} J(u) = \frac{1}{2}\|y_u(T) - y_d\|^2_{L^2(\Omega)}. \tag{$P_\alpha$}$$

Here $y_d \in L^2(\Omega)$, and $U_\alpha := \{u \in \mathcal{M}(\bar\Omega) : \|u\|_{\mathcal{M}(\bar\Omega)} \le \alpha\}$, where $\mathcal{M}(\bar\Omega)$ is the space of regular Borel measures on $\bar\Omega$ equipped with the norm

$$\|u\|_{\mathcal{M}(\bar\Omega)} := \sup_{\|\phi\|_{C(\bar\Omega)}\le 1} \int_{\bar\Omega} \phi(x)\, du(x) = |u|(\bar\Omega).$$

**Acknowledgement:** We acknowledge fruitful discussions with Eduardo Casas and Karl Kunisch, which inspired this work.

**Evelyn Herberg,** George Mason University, Department of Mathematical Sciences, 4400 University Drive, 22030 Fairfax, VA, U.S.A., e-mail: eherberg@gmu.edu
**Michael Hinze,** Universität Koblenz-Landau, Mathematisches Institut, Universitätsstraße 1, 56070 Koblenz, Germany, e-mail: hinze@uni-koblenz.de

https://doi.org/10.1515/9783110695984-006

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

The state $y_u$ solves the parabolic equation

$$\begin{cases} \partial_t y_u + A y_u = f & \text{in } Q = \Omega \times (0,T), \\ y_u(x,0) = u & \text{in } \bar{\Omega}, \\ \partial_n y_u(x,t) = 0 & \text{on } \Sigma = \Gamma \times (0,T), \end{cases} \tag{6.1}$$

where $f \in L^1(0,T;L^2(\Omega))$ is given, $\Omega \subset \mathbb{R}^n$ ($n = 1,2,3$) is an open, connected, and bounded set with Lipschitz boundary $\Gamma$, and $A$ is the elliptic operator defined by

$$A y_u := -a\Delta y_u + b(x,t) \cdot \nabla y_u + c(x,t) y_u \tag{6.2}$$

with constant $a > 0$ and functions $b \in L^\infty(Q)^n$ and $c \in L^\infty(Q)$.

The state is supposed to solve (6.1) in the following very weak sense (see, e. g., [6, Definition 2.1]):

**Definition 6.1.1.** We say that a function $y \in L^1(Q)$ is a solution of (6.1) if the following identity holds:

$$\int_Q (-\partial_t \phi + A^* \phi) y \, dxdt = \int_Q f\phi \, dxdt + \int_{\bar{\Omega}} \phi(0) \, du \quad \forall \, \phi \in \Phi, \tag{6.3}$$

where

$$\Phi := \{\phi \in L^2(0,T;H^1(\Omega)) : -\partial_t \phi + A^* \phi \in L^\infty(Q), \partial_n \phi = 0 \text{ on } \Sigma, \phi(T) = 0 \in \Omega\},$$

and $A^* \bar{\varphi} := -a\Delta\bar{\varphi} - \operatorname{div}[b(x,t)\bar{\varphi}] + c\bar{\varphi}$ is the adjoint operator of $A$.

The existence and uniqueness of solutions to the state equation (6.1) and problem ($P_\alpha$) are established in [6, Theorems 2.2 and 2.4].

Optimal control with a bound on the total variation norm of the measure-control is inspired by applications that aim at identifying pollution sources; see, e. g., [10, 19]. These problems inherit a sparsity structure (see, e. g., [11, 12, 21]), which we can retain in practical implementation by applying variational discretization, from [15] with a suitable Petrov–Galerkin approximation of the state equation (6.3); compare [14].

Let us briefly comment on related contributions in the literature. In [14] the variational discrete approach is applied to an optimal control problem with parabolic partial differential equation and space-time measure control from [5]. Control of elliptic partial differential equations with measure controls is considered in [3, 8, 9, 20], and control of parabolic partial differential equations with measure controls can be found in [4, 5, 7, 17, 18]. The novelty of the problem discussed in this work lies in constraining the control set, instead of incorporating a penalty term for the control in the target functional.

The plan of the paper is as follows: We analyze the continuous problem, its sparsity structure, and the particular case of positive controls in Section 6.2. Thereafter

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

we apply variational discretization to the optimal control problem in Section 6.3. Finally, in Section 6.4, we apply the semismooth Newton method to the optimal control problem with positive controls (Subsection 6.4.1) and to the original optimal control problem (Subsection 6.4.2). For the latter, we add a penalty term before applying the semismooth Newton method. For both cases, we provide numerical examples.

## 6.2 Continuous optimality system

In this section, we summarize properties of $(P_\alpha)$ established in [6].

Let $\bar{u}$ be the unique solution of $(P_\alpha)$ with associated state $\bar{y}$. We then say that $\bar{\varphi} \in L^2(0,T;H^1(\Omega)) \cap C(\bar{\Omega}\times[0,T])$ is the associated adjoint state of $\bar{u}$ if it solves

$$\begin{cases} -\partial_t\bar{\varphi} + A^*\bar{\varphi} = 0 & \text{in } Q,\\ \bar{\varphi}(x,T) = \bar{y}(x,T) - y_d & \text{in } \Omega,\\ \partial_n\bar{\varphi}(x,t) = 0 & \text{on } \Sigma. \end{cases} \tag{6.4}$$

We recall the optimality conditions for $(P_\alpha)$ from [6, Theorem 2.5].

**Theorem 6.2.1.** *Let $\bar{u}$ be a solution of $(P_\alpha)$ with associated state $\bar{y}$ and adjoint state $\bar{\varphi}$. Then we have the following properties:*

(1) *If $\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} < \alpha$, then $\bar{y}(T) = y_d$ and $\bar{\varphi} = 0 \in Q$.*

(2) *If $\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} = \alpha$, then*

$$\operatorname{supp}(\bar{u}^+) \subset \{x\in\bar{\Omega} : \bar{\varphi}(x,0) = -\|\bar{\varphi}(0)\|_{C(\bar{\Omega})}\},$$
$$\operatorname{supp}(\bar{u}^-) \subset \{x\in\bar{\Omega} : \bar{\varphi}(x,0) = +\|\bar{\varphi}(0)\|_{C(\bar{\Omega})}\},$$

*where $\bar{u} = \bar{u}^+ - \bar{u}^-$ is the Jordan decomposition of $\bar{u}$.*

*Conversely, if $\bar{u}$ is an element of $U_\alpha$ satisfying* (1) *or* (2)*, then $\bar{u}$ is a solution to $(P_\alpha)$.*

In some applications, we may have a priori knowledge about the measure controls. This motivates the restriction of the admissible control set to positive controls $U_\alpha^+ := \{u\in\mathcal{M}^+(\bar{\Omega}) : \|u\|_{\mathcal{M}(\bar{\Omega})} \le \alpha\}$ with $\|u\|_{\mathcal{M}(\bar{\Omega})} = u(\bar{\Omega})$. We then consider the problem

$$\min_{u\in U_\alpha^+} J(u) = \frac{1}{2}\|y_u(T) - y_d\|^2_{L^2(\Omega)}, \tag{$P_\alpha^+$}$$

where $y_u$ solves (6.1). The properties of $(P_\alpha^+)$ are derived in [6, Theorem 3.1]:

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Theorem 6.2.2.** *Problem $(P_\alpha^+)$ has a unique solution. Let $\bar{u}$ be the unique solution of $(P_\alpha^+)$ with associated adjoint state $\bar{\varphi}$. Then $\bar{u}$ is a solution of $(P_\alpha^+)$ if and only if*

$$\int_{\bar{\Omega}} \bar{\varphi}(x,0)\, d\bar{u} \le \int_{\bar{\Omega}} \bar{\varphi}(x,0)\, du \quad \forall\, u \in U_\alpha^+. \tag{6.5}$$

*If $u(\bar{\Omega}) = \alpha$, then we have the following properties:*
(1) *Inequality* (6.5) *is equivalent to the identity*

$$\int_{\bar{\Omega}} \bar{\varphi}(x,0)\, d\bar{u} = \alpha\bar{\lambda} := \alpha \min_{x\in\bar{\Omega}} \bar{\varphi}(x,0), \tag{6.6}$$

*where $\bar{\lambda} \le 0$.*
(2) *A control $\bar{u}$ is a solution of $(P_\alpha^+)$ if and only if*

$$\operatorname{supp}(\bar{u}) \subset \{x \in \bar{\Omega} : \bar{\varphi}(x,0) = \bar{\lambda}\}. \tag{6.7}$$

We also repeat the following remark from [6, Remark 3.3].

**Remark 6.2.3.** Although in Theorem 6.2.1, we have $\bar{y}(T) = y_d$ and $\bar{\varphi} = 0 \in Q$ for an optimal control $\bar{u}$ with $\bar{u}(\bar{\Omega}) < \alpha$, this case is not a part of Theorem 6.2.2. For nonnegative controls, we can show that if $y_d \le y_0(T)$, where $y_0$ is the solution of (6.1) corresponding to the control $u = 0$, then the unique solution to $(P_\alpha^+)$ is given by $\bar{u} = 0$. So even though $\bar{u}(\bar{\Omega}) = 0 < \alpha$, we have $\bar{y}(T) \ne y_d$, and, consequently, $\bar{\varphi} \ne 0 \in Q$.

## 6.3 Variational discretization

To discretize problems $(P_\alpha)$ and $(P_\alpha^+)$, we define the space-time grid as follows: Define the partition $0 = t_0 < t_1 < \cdots < t_{N_\tau} = T$. For the temporal grid, the interval $I$ is split into subintervals $I_k = (t_{k-1}, t_k]$ for $k = 1, \dots, N_\tau$. The temporal gridsize is denoted by $\tau = \max_{0\le k\le N_\tau} \tau_k$, where $\tau_k := t_k - t_{k-1}$. We assume that $\{I_k\}_k$ and $\{\mathcal{K}_h\}_h$ are quasi-uniform sequences of time grids and triangulations, respectively. For $K \in \mathcal{K}_h$, we denote by $\rho(K)$ the diameter of $K$, and $h := \max_{K\in\mathcal{K}_h} \rho(K)$. We set $\bar{\Omega}_h = \bigcup_{K\in\mathcal{K}_h} K$ and denote by $\Omega_h$ the interior and by $\Gamma_h$ the boundary of $\bar{\Omega}_h$. We assume that vertices on $\Gamma_h$ are points on $\Gamma$. We then set up the space-time grid as $Q_h := \Omega_h \times (0, T)$.

We define the discrete spaces

$$Y_h := \operatorname{span}\{\phi_j : 1 \le j \le N_h\}, \tag{6.8}$$

$$Y_\sigma := \operatorname{span}\{\phi_j \otimes \chi_k : 1 \le j \le N_h, 1 \le k \le N_\tau\}, \tag{6.9}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

where $\chi_k$ is the indicator function of $I_k$, and $(\phi_j)_{j=1}^{N_h}$ is the nodal basis formed by continuous piecewise linear functions satisfying $\phi_j(x_i) = \delta_{ij}$.

We choose the space $Y_\sigma$ as our discrete state and test space in a $dG(0)$ approximation of (6.1). The control space remains either $U_\alpha$ or $U_\alpha^+$.

This approximation scheme is equivalent to an implicit Euler time stepping scheme. To see this, we recall that the elements $y_\sigma \in Y_\sigma$ can be represented as

$$y_\sigma = \sum_{k=1}^{N_\tau} y_{k,h} \otimes \chi_k$$

with $y_{k,h} := y_\sigma|_{I_k} \in Y_h$.

Given a control $u \in U_\alpha$ for $k = 1, \dots, N_\tau$ and $z_h \in Y_h$, we thus end up with the variational discrete scheme

$$\begin{cases} (y_{k,h} - y_{k-1,h}, z_h)_{L^2} + a\,\tau_k \int_\Omega \nabla y_{k,h} \nabla z_h \, dx \\ \quad + \int_{I_k} \int_\Omega b(x,t) \nabla y_{k,h} z_h + c(x,t) y_{k,h} z_h \, dx \, dt = \int_{I_k} \int_\Omega f z_h \, dx \, dt, \\ y_{0,h} = y_{0h}, \end{cases} \tag{6.10}$$

where $y_{0h} \in Y_h$ is the unique element satisfying:

$$(y_{0h}, z_h) = \int_\Omega z_h \, du \quad \forall z_h \in Y_h. \tag{6.11}$$

Here $(\cdot,\cdot)_{L^2}$ denotes the $L^2(\Omega)$ inner product. We assume that the discretization parameters $h$ and $\tau$ are sufficiently small, such that there exists a unique solution to (6.10) for general functions $b$ and $c$.

The variational discrete counterparts to $(P_\alpha)$ and $(P_\alpha^+)$ now read

$$\min_{u \in U_\alpha} J_\sigma(u) = \frac{1}{2} \left\| y_{u,\sigma}(T) - y_d \right\|^2_{L^2(\Omega_h)} \tag{$P_{\alpha,\sigma}$}$$

and

$$\min_{u \in U_\alpha^+} J_\sigma(u) = \frac{1}{2} \left\| y_{u,\sigma}(T) - y_d \right\|^2_{L^2(\Omega_h)}, \tag{$P_{\alpha,\sigma}^+$}$$

respectively, where in both cases, $y_{u,\sigma}$ for given $u$ denotes the unique solution of (6.10). It is now straightforward to show that the optimality conditions for problems $(P_{\alpha,\sigma})$ and $(P_{\alpha,\sigma}^+)$ read like those for $(P_\alpha)$ and $(P_\alpha^+)$ with the adjoint $\varphi$ replaced by $\varphi_{\bar{u},\sigma} \in Y_h$ for given solution $\bar{u}$, the solution to the following system for $k = 1, \dots, N_\tau$ and $z_h \in Y_h$:

$$\begin{cases} -(\varphi_{k,h} - \varphi_{k-1,h}, z_h)_{L^2} + a\,\tau_k \int_\Omega \nabla \varphi_{k-1,h} \nabla z_h \, dx \\ \quad + \int_{I_k} \int_\Omega - \operatorname{div}(b(x,t) \varphi_{k-1,h}) z_h + c(x,t) \varphi_{k-1,h} z_h \, dx \, dt = 0, \\ \varphi_{N_\tau,h} = \varphi_{N_\tau h}, \end{cases} \tag{6.12}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

where $z_h \in Y_h$, and $\varphi_{N_\tau h} \in Y_h$ is the unique element satisfying

$$(\varphi_{N_\tau h}, z_h) = \int_\Omega (y_{\bar{u},\sigma}(T) - y_d) z_h \, dx \quad \forall\, z_h \in Y_h. \tag{6.13}$$

For details on the derivation of the optimality conditions, we refer to Theorems 6.3.7 and 6.3.8, which will be proven after introducing a few helpful results.

This in particular implies that

$$\begin{aligned} \operatorname{supp}(\bar{u}^+) &\subset \{x \in \bar{\Omega} : \varphi_{\bar{u},\sigma}(x,0) = -\|\varphi_{\bar{u},\sigma}(0)\|_\infty\}, \\ \operatorname{supp}(\bar{u}^-) &\subset \{x \in \bar{\Omega} : \varphi_{\bar{u},\sigma}(x,0) = +\|\varphi_{\bar{u},\sigma}(0)\|_\infty\}. \end{aligned} \tag{6.14}$$

Analogously for $(P^+_{\alpha,\sigma})$, in the case $u(\bar{\Omega}) = \alpha$, we have the optimality condition

$$\operatorname{supp}(\bar{u}) \subset \left\{x \in \bar{\Omega} : \varphi_{\bar{u},\sigma}(x,0) = \min_{x \in \bar{\Omega}} \varphi_{\bar{u},\sigma}(x,0)\right\}.$$

Since in both cases, $\varphi_{\bar{u},\sigma}$ is a piecewise linear and continuous function, the extremal value in the generic case can only be attained at grid points, which leads to

$$\operatorname{supp}(\bar{u}) \subset \{x_j\}_{j=1}^{N_h}.$$

So we derive the implicit discrete structure

$$\bar{u} \in U_h := \operatorname{span}\{\delta_{x_j} : 1 \le j \le N_h\},$$

where $\delta_{x_j}$ denotes a Dirac measure at gridpoint $x_j$. In the case of $(P^+_{\alpha,\sigma})$, we even know that all coefficients will be positive, and hence we get

$$\bar{u} \in U_h^+ := \left\{\sum_{j=1}^{N_h} u_j \delta_{x_j} : u_j \ge 0\right\}.$$

Notice also that the natural pairing $\mathcal{M}(\bar{\Omega}) \times \mathcal{C}(\bar{\Omega}) \to \mathbb{R}$ induces the duality $Y_h^* \cong U_h$ in the discrete setting. Here we see the effect of the variational discretization concept: The choice for the discretization of the test space induces a natural discretization for the controls.

We note that the use of piecewise linear and continuous Ansatz- and test-functions in the variational discretization creates a setting where the optimal control is supported on space grid points. However, it is possible to use piecewise quadratic and continuous Ansatz- and test-functions, so that the discrete adjoint variable can attain its extremal values not only on grid points, but anywhere. Calculating the location of these extremal values then would mean determining the potential support of the optimal control, not limited to grid points anymore.

The following operator will be useful for the discussion of solutions to $(P_{\alpha,\sigma})$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Lemma 6.3.1.** *Let the linear operator* $\Upsilon_h : \mathcal{M}(\bar{\Omega}) \to U_h \subset \mathcal{M}(\bar{\Omega})$ *be defined as*

$$\Upsilon_h u := \sum_{j=1}^{N_h} \delta_{x_j} \int_\Omega \phi_j \, du.$$

*Then for all* $u \in \mathcal{M}(\bar{\Omega})$ *and* $\varphi_h \in Y_h$*, we have the following properties:*

$$\langle u, \varphi_h \rangle = \langle \Upsilon_h u, \varphi_h \rangle, \tag{6.15}$$

$$\|\Upsilon_h u\|_{\mathcal{M}(\bar{\Omega})} \leq \|u\|_{\mathcal{M}(\bar{\Omega})}. \tag{6.16}$$

These results are proven in [5, Proposition 4.1]. Furthermore, it is obvious that for piecewise linear and continuous finite elements, $\Upsilon_h(\mathcal{M}^+(\bar{\Omega})) \subset U_h^+$.

The mapping $u \mapsto y_{u,\sigma}(T)$ is in general not injective, and hence the uniqueness of the solution cannot be concluded. However, in the implicitly discrete setting, we can prove uniqueness similarly as done in [4, Section 4.3] and [14, Theorem 11].

**Theorem 6.3.2.** *Problem* $(P_{\alpha,\sigma})$ *has at least one solution in* $\mathcal{M}(\bar{\Omega})$*, and there exists a unique solution* $\bar{u} \in U_h$*. Furthermore, for every solution* $\hat{u} \in \mathcal{M}(\bar{\Omega})$ *of* $(P_{\alpha,\sigma})$*, we have* $\Upsilon_h \hat{u} = \bar{u}$*. Moreover, if* $\bar{\varphi}_h(x_j) \neq \bar{\varphi}_h(x_k)$ *for all neighboring finite element nodes* $x_j \neq x_k$ *of the finite element nodes* $x_j$ $(j = 1, \dots, N_h)$*, problem* $(P_{\alpha,\sigma})$ *admits a unique solution, which is an element of* $U_h$*.*

*Proof.* The existence of solutions can be derived as for the continuous problem (see [6, Theorem 2.4]), since the control domain remains continuous. We include the details for the convenience of the reader.

The control domain $U_\alpha$ is bounded and weakly-* closed in $\mathcal{M}(\bar{\Omega})$. From the Banach–Alaoglu–Bourbaki theorem we even know that it is weakly-* compact; see, e. g., [2, Theorem 3.16]. Hence any minimizing sequence is bounded in $\mathcal{M}(\bar{\Omega})$, and any weak-* limit belongs to the control domain $U_\alpha$. Using convergence properties from [6, Theorem 2.3], we can conclude that any of these limits is a solution to $(P_{\alpha,\sigma})$.

Let $\hat{u} \in \mathcal{M}(\bar{\Omega})$ be a solution of $(P_{\alpha,\sigma})$, and let $\bar{u} := \Upsilon_h \hat{u} \in U_h$. From (6.15) we have

$$y_{u,\sigma} = y_{\Upsilon_h u,\sigma} \quad \text{for all } u \in \mathcal{M}(\bar{\Omega}).$$

From this we deduce $J_\sigma(\bar{u}) = J_\sigma(\hat{u})$. Moreover, (6.16) delivers

$$\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} \leq \|\hat{u}\|_{\mathcal{M}(\bar{\Omega})},$$

so $\bar{u}$ is admissible, since $\hat{u} \in U_\alpha$. Altogether, this shows the existence of solutions in the discrete space $U_h$.

Since the mapping $u \mapsto y_{u,\sigma}(T)$ is injective for $u \in U_h$ (since $\dim(U_h) = \dim(Y_h)$, and $J_\sigma(u)$ is a quadratic function), we deduce the strict convexity of $J_\sigma(u)$ on $U_h$. Furthermore, $\{u \in U_h : \|u\|_{\mathcal{M}(\bar{\Omega})} = \sum_{j=1}^{N_h} |u_j| \leq \alpha\}$ is a closed convex set, so we can conclude the uniqueness of the solution in the discrete space.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

For every solution $\hat{u} \in \mathcal{M}(\bar{\Omega})$ of $(P_{\alpha,\sigma})$, the projection $\Upsilon_h \hat{u}$ is a discrete solution. Moreover, there exists only one discrete solution. So we deduce that all projections must coincide.

If now $\bar{\varphi}_h(x_j) \neq \bar{\varphi}_h(x_k)$ for all neighbors $k \neq j$, then every solution $u$ of $(P_{\alpha,\sigma})$ has its support in some of the finite element nodes of the triangulation or vanishes identically and thus is an element of $U_h$. This shows the unique solvability of $(P_{\alpha,\sigma})$ in this case. □

**Remark 6.3.3.** We note that the condition on the values of $\bar{\varphi}_h$ in the finite element nodes for guaranteeing the uniqueness can be checked once the discrete adjoint solution is known. This condition is thus fully practical.

For $(P^+_{\alpha,\sigma})$, we have a similar result like Theorem 6.3.2, which we state without proof, since it can be interpreted as a particular case of Theorem 6.3.2 and can be proven analogously.

**Theorem 6.3.4.** *Problem $(P^+_{\alpha,\sigma})$ has at least one solution in $\mathcal{M}^+(\bar{\Omega})$, and there exists a unique solution $\bar{u} \in U_h^+$. Furthermore, for every solution $\hat{u} \in \mathcal{M}^+(\bar{\Omega})$ of $(P^+_{\alpha,\sigma})$, we have $\Upsilon_h \hat{u} = \bar{u}$. Moreover, if $\bar{\varphi}_h(x_j) \neq \bar{\varphi}_h(x_k)$ for all neighboring finite element nodes $x_j \neq x_k$ of the finite element nodes $x_j$ $(j = 1, \dots, N_h)$, problem $(P^+_{\alpha,\sigma})$ admits a unique solution, which is an element of $U_h^+$.*

Now we introduce two useful lemmas.

**Lemma 6.3.5.** *Given $u \in \mathcal{M}(\bar{\Omega})$, the solution $z_{u,\sigma} \in Y_h$ to (6.10) with $f \equiv 0$ satisfies*

$$\int_\Omega (y_{u,\sigma}(T) - y_d) z_{u,\sigma}(T)\, dx = \int_\Omega \varphi_{u,\sigma}(0)\, du. \tag{6.17}$$

*Proof.* We take (6.10) with $f \equiv 0$ and test with the components $\varphi_{k,h}$ of $\varphi_{u,\sigma}$ for all $k = 1, \dots, N_\tau$. Similarlym we take (6.12) and test this with the components $z_{k-1,h}$ of $z_{u,\sigma}$ for all $k = 1, \dots, N_\tau$. Now we can sum up the equations, and since in both cases the right-hand side is zero, we can equalize those sums. Furthermore, we can apply Gauss' theorem and drop all terms that appear on both sides. This leads to

$$\begin{aligned}
&\sum_{k=1}^{N_\tau} (z_{k,h} - z_{k-1,h}, \varphi_{k,h}) = \sum_{k=1}^{N_\tau} (-\varphi_{k,h} + \varphi_{k-1,h}, z_{k-1,h}) \\
\Rightarrow\quad &\sum_{k=1}^{N_\tau} (z_{k,h}, \varphi_{k,h}) - (z_{k-1,h}, \varphi_{k,h}) = \sum_{k=1}^{N_\tau} -(\varphi_{k,h}, z_{k-1,h}) + (\varphi_{k-1,h}, z_{k-1,h}) \\
\Rightarrow\quad &\sum_{k=1}^{N_\tau} (z_{k,h}, \varphi_{k,h}) = \sum_{k=0}^{N_\tau - 1} (\varphi_{k,h}, z_{k,h}) \\
\Rightarrow\quad &(z_{N_\tau,h}, \varphi_{N_\tau,h}) = (\varphi_{0,h}, z_{0,h}).
\end{aligned}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

We have $z_{N_\tau,h} = z_{u,\sigma}(T) \in Y_h$ and $\varphi_{0,h} = \varphi_{u,\sigma}(0) \in Y_h$, so together with (6.11) and (6.13), we can deduce (6.17). □

**Lemma 6.3.6.** *For every $\epsilon > 0$ and all $h$ small enough, there exists a control $u \in L^2(\Omega)$ such that the solution $y_{u,\sigma}$ of* (6.10) *fulfills*

$$\|y_{u,\sigma}(T) - y_d\|_{L^2(\Omega_h)} < \epsilon. \tag{6.18}$$

*Proof.* Let $y_{d,\sigma}$ be the $L^2$-projection of $y_d$ onto $Y_h$. Then for $h$ small enough,

$$\|y_{u,\sigma}(T) - y_d\|_{L^2(\Omega_h)} \le \|y_{u,\sigma}(T) - y_{d,\sigma}\|_{L^2(\Omega_h)} + \underbrace{\|y_{d,\sigma} - y_d\|_{L^2(\Omega_h)}}_{<\epsilon}.$$

Let additionally $\tau$ be small enough such that scheme (6.10) has a unique solution. Then in every time step, we obtain a system of equations, where the matrix is an isomorphism on $Y_h$. Consequently, the initial to final value map $y_{0h} \mapsto y_{u,\sigma}(T)$ is an isomorphism. Since $Y_h \subset L^2(\Omega_h)$, we can find $u \in L^2(\Omega)$ such that

$$\|y_{u,\sigma}(T) - y_{d,\sigma}\|_{L^2(\Omega_h)} = 0,$$

which completes the proof. □

Finally, we give discrete versions of Theorems 6.2.1 and 6.2.2. Both are proven very similarly to the continuous case; see [6, Theorems 2.5 and 3.1].

**Theorem 6.3.7.** *Let $\bar{u}$ solve $(P_{\alpha,\sigma})$ with the associated discrete state $y_{\bar{u},\sigma}$ and discrete adjoint state $\varphi_{\bar{u},\sigma}$. Then for $\sigma$ small enough:*
(1) *if $\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} < \alpha$, then $y_{\bar{u},\sigma}(T) = y_d$ and $\varphi_{\bar{u},\sigma} = 0 \in Q$;*
(2) *if $\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} = \alpha$, then*

$$\operatorname{supp}(\bar{u}^+) \subset \{x \in \bar{\Omega} : \varphi_{\bar{u},\sigma}(x,0) = -\|\varphi_{\bar{u},\sigma}(0)\|_{C(\bar{\Omega})}\}, \tag{6.19}$$

$$\operatorname{supp}(\bar{u}^-) \subset \{x \in \bar{\Omega} : \varphi_{\bar{u},\sigma}(x,0) = +\|\varphi_{\bar{u},\sigma}(0)\|_{C(\bar{\Omega})}\}, \tag{6.20}$$

*where $\bar{u} = \bar{u}^+ - \bar{u}^-$ is the Jordan decomposition of $\bar{u}$.*

*Conversely, if $\bar{u}$ is an element of $U_\alpha$ satisfying* (1) *or* (2)*, then $\bar{u}$ is a solution to $(P_{\alpha,\sigma})$.*

*Proof.* Let $u \in U_\alpha$ be arbitrary and denote by $z_{(u-\bar{u}),\sigma}$ the solution to (6.10) with $f \equiv 0$ and $u$ replaced by $u - \bar{u}$. From Lemma 6.3.5 we get

$$\begin{aligned}\lim_{\rho \searrow 0} \frac{J_\sigma(\bar{u} + \rho(u - \bar{u})) - J_\sigma(\bar{u})}{\rho} &= \int_\Omega (y_{\bar{u},\sigma}(T) - y_d) z_{(u-\bar{u}),\sigma}(T)\, dx \\ &= \int_\Omega \varphi_{\bar{u},\sigma}(0)\, d(u - \bar{u}).\end{aligned}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Since $(P_{\alpha,\sigma})$ is a convex problem, the following variational inequality is a necessary and sufficient condition for optimality of a control $\bar{u} \in U_\alpha$:

$$\begin{aligned} J'_\sigma(\bar{u})(u-\bar{u}) = \int_\Omega \varphi_{\bar{u},\sigma}(0)\, d(u-\bar{u}) \geq 0 \quad \forall u \in U_\alpha \\ \Rightarrow \quad -\int_\Omega \varphi_{\bar{u},\sigma}(0)\, du \leq -\int_\Omega \varphi_{\bar{u},\sigma}(0)\, d\bar{u} \quad \forall u \in U_\alpha \\ \Rightarrow \quad \sup_{u\in U_\alpha} \int_\Omega \varphi_{\bar{u},\sigma}(0)\, du = -\int_\Omega \varphi_{\bar{u},\sigma}(0)\, d\bar{u} \\ \Rightarrow \quad \alpha \|\varphi_{\bar{u},\sigma}(0)\|_{\mathcal{C}(\bar{\Omega})} = -\int_\Omega \varphi_{\bar{u},\sigma}(0)\, d\bar{u}. \end{aligned}$$

Let $\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} = \alpha$. Then this is

$$\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} \|\varphi_{\bar{u},\sigma}(0)\|_{\mathcal{C}(\bar{\Omega})} = -\int_\Omega \varphi_{\bar{u},\sigma}(0)\, d\bar{u}. \tag{6.21}$$

We now may conclude as in [4, Lemma 3.4] to obtain (6.19) and (6.20). Also, if these conditions hold, then we get equality (6.21), which is a necessary and sufficient condition for optimality of $\bar{u}$; so $\bar{u}$ solves $(P_{\alpha,\sigma})$.

Let us now study the case $\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} < \alpha$. If $y_{\bar{u},\sigma} = y_d$, then $J_\sigma(\bar{u}) = 0$, and since $J_\sigma(u) \geq 0$ for all $u \in U_\alpha$, we deduce that $\bar{u}$ solves $(P^+_{\alpha,\sigma})$. Now assume that $\bar{u}$ solves $(P^+_{\alpha,\sigma})$ and $y_{\bar{u},\sigma} \neq y_d$ holds. Then we have $J_\sigma(\bar{u}) > 0$. From Lemma 6.3.6 we know that for $h$ small enough, there exists an element $u \in \mathcal{M}(\bar{\Omega})$ such that $J_\sigma(u) < J_\sigma(\bar{u})$. Since $\bar{u}$ is a solution to $(P^+_{\alpha,\sigma})$, we have $u \notin U_\alpha$. Now take $\lambda \in \mathbb{R}$ such that

$$0 < \lambda < \min\left\{ \frac{\alpha - \|\bar{u}\|_{\mathcal{M}(\bar{\Omega})}}{\|u-\bar{u}\|_{\mathcal{M}(\bar{\Omega})}}, 1 \right\}. \tag{6.22}$$

Then $v := \bar{u} + \lambda(u-\bar{u}) \in U_\alpha$, and by the convexity of $J_\sigma$ we get

$$J_\sigma(v) = J_\sigma\big(\lambda u + (1-\lambda)\bar{u}\big) \leq \lambda \underbrace{J_\sigma(u)}_{<J_\sigma(\bar{u})} + (1-\lambda) J_\sigma(\bar{u}) < J_\sigma(\bar{u}),$$

so that $\bar{u} \in U_\alpha$ cannot be a solution of $(P^+_{\alpha,\sigma})$. Hence $y_{\bar{u},\sigma} = y_d$, and from (6.13) we deduce $\varphi_{\bar{u},\sigma} = 0$. □

**Theorem 6.3.8.** *Let $\bar{u}$ solve $(P^+_{\alpha,\sigma})$ with associated discrete adjoint state $\varphi_{\bar{u},\sigma}$. Then $\bar{u}$ is a solution of $(P^+_{\alpha,\sigma})$ if and only if*

$$\int_\Omega \varphi_{\bar{u},\sigma}(x,0)\, d\bar{u} \leq \int_\Omega \varphi_{\bar{u},\sigma}(x,0)\, du \quad \forall\, u \in U^+_\alpha. \tag{6.23}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*If $\bar{u}(\bar{\Omega}) = \alpha$, then we have the following properties:*
(1) *inequality* (6.23) *is equivalent to the identity*

$$\int_{\Omega} \varphi_{\bar{u},\sigma}(x,0)\, d\bar{u} = \alpha\bar{\lambda} := \alpha \min_{x\in\bar{\Omega}} \varphi_{\bar{u},\sigma}(x,0), \tag{6.24}$$

*where $\bar{\lambda} \le 0$;*
(2) *$\bar{u}$ is the solution of $(P_\alpha^+)$ if and only if*

$$\operatorname{supp}(\bar{u}) \subset \{x \in \bar{\Omega} : \varphi_{\bar{u},\sigma}(x,0) = \bar{\lambda}\}. \tag{6.25}$$

*Proof.* As in the proof of Theorem 6.3.7, we get that $\bar{u} \in U_\alpha^+$ solves $(P_{\alpha,\sigma}^+)$ if and only if

$$J'_\sigma(\bar{u})(u-\bar{u}) = \int_{\Omega} \varphi_{\bar{u},\sigma}(0)\, d(u-\bar{u}) \ge 0 \quad \forall u \in U_\alpha^+,$$

which is equivalent to condition (6.23).

Now let $\bar{u}(\bar{\Omega}) = \alpha$. If $\bar{\lambda} = \min_{x\in\bar{\Omega}} \varphi_{\bar{u},\sigma}(x,0) > 0$, then take $u = 0 \in U_\alpha^+$ in (6.23) to see that $\bar{u} = 0$ in this case. So we must have $\bar{\lambda} \le 0$. Furthermore, we can equivalently write (6.23) as

$$\int_{\Omega} \varphi_{\bar{u},\sigma}(x,0)\, d\bar{u} = \min_{u\in U_\alpha^+} \int_{\Omega} \varphi_{\bar{u},\sigma}(x,0)\, du.$$

Take $x_0 \in \bar{\Omega}$ such that $\varphi_{\bar{u},\sigma}(x_0,0) = \bar{\lambda}$. Then $u = \alpha\delta_{x_0}$ achieves the minimum in the equation above, and we get (6.24). The other direction of the equivalence is obvious and completes the proof of part (1).

To prove part (2), we look at two cases. First, let $\bar{\lambda} = 0$. By the definition of $\bar{\lambda}$ this implies that $\varphi_{\bar{u},\sigma} \ge 0$ for all $x \in \bar{\Omega}$. So with (6.24) we get that $\bar{u}$ has support, where $\varphi_{\bar{u},\sigma}(x,0) = 0 = \bar{\lambda}$, in order for the integral to be zero.

The second case is $\bar{\lambda} < 0$. Define $\psi(x) := -\min\{\varphi_{\bar{u},\sigma}(x,0), 0\}$, then $0 \le \psi(x) \le -\bar{\lambda}$ by the definition of $\psi(x)$ and $\bar{\lambda}$. Furthermore, $\|\psi\|_{\mathcal{C}(\bar{\Omega})} = -\bar{\lambda}$. With (6.23) and $\psi(x) \ge -\varphi_{\bar{u},\sigma}(x,0)$, we find

$$\int_{\Omega} \psi(x)\, d\bar{u} \ge -\int_{\Omega} \varphi_{\bar{u},\sigma}(x,0)\, d\bar{u} \ge -\int_{\Omega} \varphi_{\bar{u},\sigma}(x,0)\, du \quad \forall u \in U_\alpha^+.$$

In particular, for $u = \alpha\delta_{x_0}$, we have

$$\int_{\Omega} \psi(x)\, d\bar{u} \ge -\int_{\Omega} \varphi_{\bar{u},\sigma}(x,0)\, d(\alpha\delta_{x_0}) = -\alpha\bar{\lambda} = \|\bar{u}\|_{\mathcal{M}(\bar{\Omega})}\|\psi\|_{\mathcal{C}(\bar{\Omega})}.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Furthermore, we obviously have

$$\int_{\Omega} \psi(x)\, d\bar{u} \le \|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} \|\psi\|_{\mathcal{C}(\bar{\Omega})},$$

so we can deduce the equality, and then by [4, Lemma 3.4] we get (6.25).

The converse implication can be seen, since for a positive control $\bar{u} \in U_\alpha^+$ with $\bar{u}(\bar{\Omega}) = \alpha$, we can follow (6.23) from condition (6.25). □

## 6.4 Numerical results

For the implementation, we consider $b \equiv 0$ and $c \equiv 0$ in (6.2).

We will first consider the case of positive sources, since the implementation is straightforward, whereas the general case requires to handle absolute values in the constraints.

### 6.4.1 Positive sources (problem $(P^+_{\alpha,\sigma})$)

We recall the discrete state equation (6.10), which reduces to the following form, since $b \equiv 0$ and $c \equiv 0$, with $z_h \in Y_h$:

$$\begin{cases} (y_{k,h} - y_{k-1,h}, z_h)_{L^2} + \tau_k \int_\Omega \nabla y_{k,h} \nabla z_h \, dx = \int_{I_k} \int_\Omega f\, z_h \, dx\, dt, \\ y_{0,h} = y_{0h}, \end{cases}$$

where $y_{0h} \in Y_h$, for given $u \in \mathcal{M}(\bar{\Omega})$, is the unique element satisfying

$$(y_{0h}, z_h) = \int_\Omega z_h \, du \quad \forall\, z_h \in Y_h.$$

We define the mass matrix $M_h := ((\phi_j, \phi_k)_{L^2})_{j,k=1}^{N_h}$ and the stiffness matrix $A_h := (\int_{\bar{\Omega}} \nabla\phi_j \nabla\phi_k)_{j,k=1}^{N_h}$ corresponding to $Y_h$.

We also notice that the matrix $(\phi_j, \delta_{x_k})_{j,k=1}^{N_h}$ is the identity in $\mathbb{R}^{N_h \times N_h}$. We represent the discrete state equation by the following operator $L : \mathbb{R}^{N_\sigma} \to \mathbb{R}^{N_\sigma}$:

$$\underbrace{\begin{pmatrix} M_h & & & 0 \\ -M_h & M_h + \tau_1 A_h & & \\ & \ddots & \ddots & \\ 0 & & -M_h & M_h + \tau_{N_\tau} A_h \end{pmatrix}}_{=:L} \begin{pmatrix} y_{0,h} \\ y_{1,h} \\ \vdots \\ y_{N_\tau,h} \end{pmatrix} = \begin{pmatrix} u \\ \tau_1 M_h f_{1,h} \\ \vdots \\ \tau_{N_\tau} M_h f_{N_\tau,h} \end{pmatrix}. \tag{6.26}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

We can now formulate the following finite-dimensional discrete problem $(P^+_{\alpha,\sigma})$:

$$\begin{aligned} \min_{u\in\mathbb{R}^{N_h}} J(u) &= \frac{1}{2}\big(y_{N_\tau,h}(u)-y_d\big)^\top M_h\big(y_{N_\tau,h}(u)-y_d\big) \qquad (P^+_h)\\ \text{s.t.}\quad & \sum_{i=1}^{N_h} u_i - \alpha \le 0,\\ & -u_i \le 0 \quad \forall\, i\in\{1,\dots,N_h\}. \end{aligned}$$

The corresponding Lagrangian function $\mathcal{L}(u,\mu^{(1)},\mu^{(2)})$ with $\mu^{(1)}\in\mathbb{R}$ and $\mu^{(2)}\in\mathbb{R}^{N_h}$ is defined by

$$\mathcal{L}(u,\mu^{(1)},\mu^{(2)}) := J(u) + \mu^{(1)}\left(\sum_{i=1}^{N_h} u_i - \alpha\right) - \sum_{i=1}^{N_h}\mu_i^{(2)}u_i.$$

All inequalities in $(P^+_h)$ are strictly fulfilled for all $u_i = \frac{\alpha}{N_h+1}$, $i\in\{1,\dots,N_h\}$, thus an interior point of the feasible set exists, and the Slater condition is satisfied (see, e. g., [16, (1.132)]). Then the Karush–Kuhn–Tucker conditions (see, e. g., [1, (5.49)]) state that at the minimum $u$ the following conditions hold:

(1) $\partial_u\mathcal{L}(u,\mu^{(1)},\mu^{(2)}) = 0$,
(2) $\mu^{(1)}(\sum_{i=1}^{N_h} u_i-\alpha) = 0 \ \wedge\ \mu^{(1)}\ge 0 \ \wedge\ (\sum_{i=1}^{N_h} u_i-\alpha)\le 0$,
(3) $-\mu_i^{(2)}u_i = 0 \ \wedge\ \mu_i^{(2)}\ge 0 \ \wedge\ -u_i\le 0 \quad \forall\, i\in\{1,\dots,N_h\}$,

where (2) and (3) can be equivalently reformulated with arbitrary $\kappa>0$ by

$$\begin{aligned} N^{(1)}(u,\mu^{(1)}) &:= \max\left\{0,\mu^{(1)}+\kappa\left(\sum_{i=1}^{N_h}u_i-\alpha\right)\right\}-\mu^{(1)} = 0,\\ N^{(2)}(u,\mu^{(2)}) &:= \max\{0,\mu^{(2)}-\kappa u\}-\mu^{(2)} = 0. \end{aligned}$$

We define

$$F(u,\mu^{(1)},\mu^{(2)}) := \begin{pmatrix}\partial_u\mathcal{L}(u,\mu^{(1)},\mu^{(2)}) & N^{(1)}(u,\mu^{(1)}) & N^{(2)}(u,\mu^{(2)})\end{pmatrix}^\top$$

and apply the semismooth Newton method to solve $F(u,\mu^{(1)},\mu^{(2)})=0$. We have

$$\partial_u\mathcal{L}(u,\mu^{(1)},\mu^{(2)}) = \partial_u J(u) + \mu^{(1)}\mathbb{1}_{N_h} - \mu^{(2)}.$$

When setting up the matrix $DF = DF(u,\mu^{(1)},\mu^{(2)})$, we always choose $\partial_x(\max\{0,g(x)\}) = \partial_x g(x)$ if $g(x)=0$. This delivers

$$DF := \begin{pmatrix} \partial_u^2\mathcal{L}(u,\mu^{(1)},\mu^{(2)}) & \partial_{\mu^{(1)}}(\partial_u\mathcal{L}(u,\mu^{(1)},\mu^{(2)})) & \partial_{\mu^{(2)}}(\partial_u\mathcal{L}(u,\mu^{(1)},\mu^{(2)}))\\ \partial_u N^{(1)}(u,\mu^{(1)}) & \partial_{\mu^{(1)}}N^{(1)}(u,\mu^{(1)}) & 0\\ \partial_u N^{(2)}(u,\mu^{(2)}) & 0 & \partial_{\mu^{(2)}}N^{(2)}(u,\mu^{(2)}) \end{pmatrix}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

with entries

$$\begin{aligned}
\partial_u^2 \mathcal{L}(u,\mu^{(1)},\mu^{(2)}) &= \partial_u^2 J(u),\\
\partial_{\mu^{(1)}}(\partial_u \mathcal{L}(u,\mu^{(1)},\mu^{(2)})) &= \mathbb{1}_{N_h},\\
\partial_{\mu^{(2)}}(\partial_u \mathcal{L}(u,\mu^{(1)},\mu^{(2)})) &= -\mathbb{1}_{N_h\times N_h},\\
\partial_u N^{(1)}(u,\mu^{(1)}) &= \begin{cases} \kappa\,\mathbb{1}_{N_h}^\top, & \mu^{(1)} + \kappa(\sum_{i=1}^{N_h} u_i - \alpha) \geq 0,\\ 0, & \text{otherwise,}\end{cases}\\
\partial_{\mu^{(1)}} N^{(1)}(u,\mu^{(1)}) &= \begin{cases} 0, & \mu^{(1)} + \kappa(\sum_{i=1}^{N_h} u_i - \alpha) \geq 0,\\ -1, & \text{otherwise,}\end{cases}\\
\partial_{u_j} N_i^{(2)}(u,\mu^{(2)}) &= \begin{cases} -\kappa\,\delta_{ij}, & \mu_i^{(2)} - \kappa u_i \geq 0,\\ 0, & \text{otherwise,}\end{cases}\\
\partial_{\mu_j^{(2)}} N_i^{(2)}(u,\mu^{(2)}) &= \begin{cases} 0, & \mu_i^{(2)} - \kappa u_i \geq 0,\\ -\delta_{ij}, & \text{otherwise.}\end{cases}
\end{aligned}$$

**Numerical example**

Let $\Omega = [0,1]$, $T = 1$, and $a = \frac{1}{100}$. We work on an equidistant $20 \times 20$ grid for this example.

To generate a desired state $y_d$, we choose $u_{\text{true}} = \delta_{0.5}$ and $f \equiv 0$, solve the state equation on a very fine grid ($1000\times1000$), and take the evaluation of the result in $t = T$ on the current grid $\Omega_h$ as the desired state $y_d$ (see Figure 6.1). Now we can insert this $y_d$ into our problem and solve for different values of $\alpha$. Knowing the true solution $u_{\text{true}}$, we can compare our results to it. We also know that $u_{\text{true}}(\Omega) = 1$ and $\operatorname{supp}(u_{\text{true}}) = \{0.5\}$. We always start the algorithm with the control being identically zero and terminate when the residual is below $10^{-15}$.

The first case we investigate is $\alpha = 0.1$ (see Figure 6.2). This $\alpha$ is smaller than the total variation of the true control, and we observe $\bar{u}(\bar{\Omega}) = \alpha$. Furthermore, $\bar{\lambda} = \min_{x\in\bar{\Omega}} \bar{\varphi}(0) \approx -35.859$, and we can verify the optimality conditions (6.24) and (6.25),

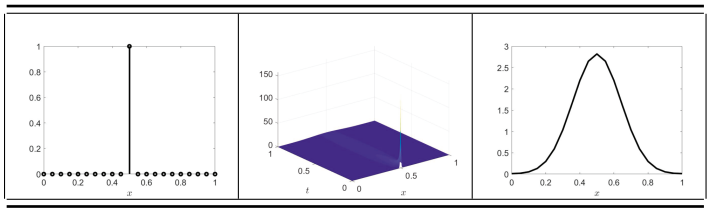


**Figure 6.1:** From left to right: the true solution $u_{\text{true}}$, associated true state $y_{\text{true}}$ in $Q = [0,1]\times[0,1]$, the desired state $y_d = y_{\text{true}}(T)$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

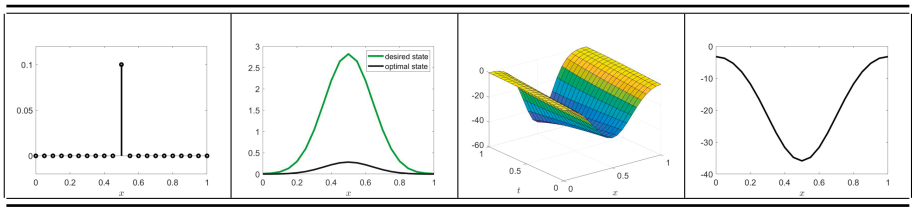



**Figure 6.2:** Solutions for $\alpha = 0.1$: from left to right: the optimal control $\bar{u}$ (solved with the semismooth Newton method), associated optimal state $\bar{y}$, associated adjoint $\bar{\varphi}$ on the whole space-time domain $Q$, and associated adjoint $\bar{\varphi}$ at $t = 0$. Terminated after 16 Newton steps.

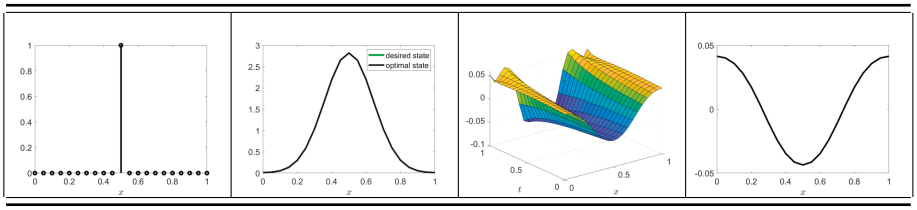



**Figure 6.3:** Solutions for $\alpha = 1$: from left to right: the optimal control $\bar{u}$ (solved with the semismooth Newton method), associated optimal state $\bar{y}$, associated adjoint $\bar{\varphi}$ on the whole space-time domain $Q$, associated adjoint $\bar{\varphi}$ at $t = 0$. Terminated after 15 Newton steps.

since

$$\int_{\bar{\Omega}} \bar{\varphi}_h(x,0)\, d\bar{u} \approx -3.5859 \approx \alpha\bar{\lambda}$$

and $\operatorname{supp}(\bar{u}) = \{0.5\}$.

The second case we investigate is $\alpha = 1 = u_{\text{true}}(\bar{\Omega})$ (see Figure 6.3). The computed optimal control in this case has a total variation of $\bar{u}(\bar{\Omega}) = 1 = \alpha$, and we can again verify the sparsity $\operatorname{supp}(\bar{u}) = \{0.5\}$. Furthermore, $\bar{\lambda} = \min_{x\in\bar{\Omega}} \bar{\varphi}(0) \approx -0.0436$, and we can verify the optimality condition (6.24), since

$$\int_{\bar{\Omega}} \bar{\varphi}_h(x,0)\, d\bar{u} \approx -0.0436 \approx \alpha\bar{\lambda}.$$

For cases with $\alpha > u_{\text{true}}(\bar{\Omega})$, we get similar results as in the case with $\alpha = 1$. In particular, this means that we observe optimality conditions (6.24) and (6.25). Since we fixed $f \equiv 0$, we get $y_0(T) \equiv 0$, and therefore $y_d > y_0(T)$. Still, the properties that we found in the general case for $\bar{u}(\bar{\Omega}) < \alpha$, $\bar{y}(T) = y_d$ and $\varphi = 0 \in Q$, cannot be observed (compare Figure 6.4, top). This is caused by the fact that the desired state $y_d$ cannot be reached on the coarse grid, so $\bar{y}(T) = y_d$ is not possible. Solving the problem with a desired state that has been projected onto the coarse grid, and thus is reachable,

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

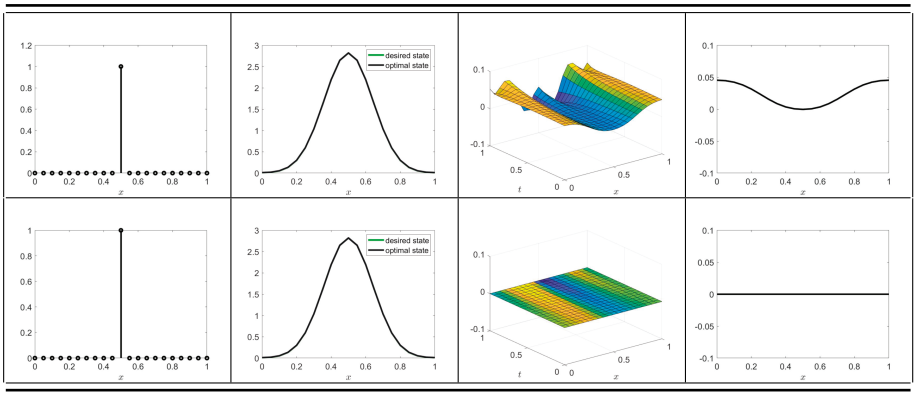



**Figure 6.4:** Solutions for $\alpha = 2$ with original desired state (top) and reachable desired state (bottom): from left to right: the optimal control $\bar{u}$ (solved with the semismooth Newton method), associated optimal state $\bar{y}$, associated adjoint $\bar{\varphi}$ on the whole space-time domain $Q$, and associated adjoint $\bar{\varphi}$ at $t = 0$. Terminated after 17 and 27 Newton steps, respectively.

delivers the expected properties $\bar{y}(T) = y_d$ and $\varphi = 0 \in Q$ (see Figure 6.4, bottom). For examples with $y_d \leq y_0(T)$, we can confirm Remark 6.2.3 and find the optimal solution $\bar{u} = 0$.

## 6.4.2 The general case (problem $(P_{\alpha,\sigma})$)

Here the source does not need to be positive.

In the discrete problem, we will decompose the control $u \in U_h$ into its positive and negative parts such that

$$u = u^+ - u^-, \quad u^+ \geq 0, \quad u^- \geq 0.$$

We have the following finite-dimensional formulation of the discrete problem $(P_{\alpha,\sigma})$:

$$\begin{aligned} \min_{u^+,u^- \in \mathbb{R}^{N_h}} \; & J(u^+,u^-) = \frac{1}{2}\left(y_{N_\tau,h}(u^+,u^-) - y_d\right)^\top M_h \left(y_{N_\tau,h}(u^+,u^-) - y_d\right) && (P_h) \\ \text{s.t.} \; & \sum_{i=1}^{N_h} \left|u_i^+ - u_i^-\right| - \alpha \leq 0, \\ & -u_i^+ \leq 0 \quad \forall\, i, \\ & -u_i^- \leq 0 \quad \forall\, i, \end{aligned}$$

where $y_{N_\tau,h}(u^+,u^-)$ corresponds to solving (6.26) with $u = u^+ - u^-$ inserted into the right-hand side of the equation. To allow taking second derivatives of the Lagrangian,

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

we want to equivalently reformulate the absolute value in the first constraint. This can be done by adding the following constraint in our discrete problem:

$$u_i^+ u_i^- = 0 \quad \forall i, \tag{6.27}$$

and, consequently, the first constraint becomes

$$\left( \sum_{i=1}^{N_h} u_i^+ + u_i^- \right) - \alpha \leq 0.$$

However, in the case $u_i^+ = u_i^- = 0$, the matrix in the Newton step will be singular. Since we want to handle sparse problems, this case will very likely occur, so we need to find a way to overcome this difficulty. Instead of adding an additional constraint, we could also add a penalty term that enforces $u_i^+ u_i^- = 0$ for all $i$ and consider the problem

$$\begin{aligned} \min_{u^+, u^- \in \mathbb{R}^{N_h}} \; & J(u^+, u^-) + \gamma (u^+)^\top u^-, \qquad (P_{h,\gamma}) \\ \text{s. t.} \quad & \sum_{i=1}^{N_h} u_i^+ + u_i^- - \alpha \leq 0, \\ & -u_i^+ \leq 0 \quad \forall i, \\ & -u_i^- \leq 0 \quad \forall i. \end{aligned}$$

For $\gamma$ large enough, the solutions of $(P_{h,\gamma})$ and $(P_h)$ will coincide. In [13, Theorem 4.6], it is specified that $\gamma$ should be larger than the largest absolute value of the Karush–Kuhn–Tucker multipliers corresponding to the equality constraints (6.27), which are replaced.

We have the corresponding Lagrangian with $\mu^{(1)} \in \mathbb{R}$ and $\mu^{(2)}, \mu^{(3)} \in \mathbb{R}^{N_h}$:

$$\begin{aligned} \mathcal{L}(u^+, u^-, \mu^{(1)}, \mu^{(2)}, \mu^{(3)}) := J(u^+, u^-) + \gamma (u^+)^\top u^- + \mu^{(1)} \left( \sum_{i=1}^{N_h} u_i^+ - u_i^- - \alpha \right) \\ - \sum_{i=1}^{N_h} \mu_i^{(2)} u_i^+ - \sum_{i=1}^{N_h} \mu_i^{(3)} u_i^-. \end{aligned}$$

All inequalities in $(P_{h,\gamma})$ are strictly fulfilled for all $u_i^+ = u_i^- = \frac{\alpha}{2(N_h+1)}$, $i \in \{1, \dots, N_h\}$, so the Slater condition is satisfied (see, e. g., [16, (1.132)]). By the Karush–Kuhn–Tucker conditions (see, e. g., [1, (5.49)]) the following conditions in the minimum $(u^+, u^-)$ must be fulfilled, where we directly reformulate the inequality conditions with arbitrary $\kappa > 0$ as in the case with positive measures:

(1) $\partial_{u^+} \mathcal{L}(u^+, u^-, \mu^{(1)}, \mu^{(2)}, \mu^{(3)}) = 0$,
(2) $\partial_{u^-} \mathcal{L}(u^+, u^-, \mu^{(1)}, \mu^{(2)}, \mu^{(3)}) = 0$,
(3) $N^{(1)}(u^+, u^-, \mu^{(1)}) = \max\{0, \mu^{(1)} + \kappa(\sum_{i=1}^{N_h} u_i^+ - u_i^- - \alpha)\} - \mu^{(1)} = 0$,

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

(4) $N^{(2)}(u^+, \mu^{(2)}) = \max\{0, \mu^{(2)} - \kappa u^+\} - \mu^{(2)} = 0,$
(5) $N^{(3)}(u^-, \mu^{(3)}) = \max\{0, \mu^{(3)} - \kappa u^-\} - \mu^{(3)} = 0.$

We then apply the semismooth Newton method to solve

$$F(u^+, u^-, \mu^{(1)}, \mu^{(2)}, \mu^{(3)}) := \begin{pmatrix} \partial_{u^+}\mathcal{L}(u^+, u^-, \mu^{(1)}, \mu^{(2)}, \mu^{(3)}) \\ \partial_{u^-}\mathcal{L}(u^+, u^-, \mu^{(1)}, \mu^{(2)}, \mu^{(3)}) \\ N^{(1)}(u^+, u^-, \mu^{(1)}) \\ N^{(2)}(u^+, \mu^{(2)}) \\ N^{(3)}(u^-, \mu^{(3)}) \end{pmatrix} = 0.$$

We have

$$\begin{aligned} \partial_{u^+}\mathcal{L}(u^+, u^-, \mu^{(1)}, \mu^{(2)}, \mu^{(3)}) &= \partial_{u^+}J(u^+, u^-) + \gamma u^- + \mu^{(1)}\mathbb{1}_{N_h} - \mu^{(2)}, \\ \partial_{u^-}\mathcal{L}(u^+, u^-, \mu^{(1)}, \mu^{(2)}, \mu^{(3)}) &= \partial_{u^-}J(u^+, u^-) + \gamma u^+ + \mu^{(1)}\mathbb{1}_{N_h} - \mu^{(3)}. \end{aligned}$$

When setting up the matrix $DF$, we always make the choice $\partial_x(\max\{0, g(x)\}) = \partial_x g(x)$ if $g(x) = 0$. This delivers (in short notation)

$$DF := \begin{pmatrix} \partial^2_{u^+}\mathcal{L} & \partial_{u^-}\partial_{u^+}\mathcal{L} & \partial_{\mu^{(1)}}\partial_{u^+}\mathcal{L} & \partial_{\mu^{(2)}}\partial_{u^+}\mathcal{L} & 0 \\ \partial_{u^+}\partial_{u^-}\mathcal{L} & \partial^2_{u^-}\mathcal{L} & \partial_{\mu^{(1)}}\partial_{u^-}\mathcal{L} & 0\partial_{\mu^{(3)}}\partial_{u^-}\mathcal{L} & \\ \partial_{u^+}N^{(1)} & \partial_{u^-}N^{(1)} & \partial_{\mu^{(1)}}N^{(1)} & 0 & 0 \\ \partial_{u^+}N^{(2)} & 0 & 0 & \partial_{\mu^{(2)}}N^{(2)} & 0 \\ 0 & \partial_{u^-}N^{(3)} & 0 & 0 & \partial_{\mu^{(3)}}N^{(3)} \end{pmatrix}$$

with the entries

$$\begin{aligned} \partial^2_{u^+}\mathcal{L} &= \partial^2_{u^+}J, \\ \partial_{u^-}\partial_{u^+}\mathcal{L} &= \partial_{u^+}\partial_{u^-}\mathcal{L} = \partial_{u^-}\partial_{u^+}J + \gamma\mathbb{1}_{N_h\times N_h}, \\ \partial_{\mu^{(1)}}\partial_{u^+}\mathcal{L} &= \partial_{\mu^{(1)}}\partial_{u^-}\mathcal{L} = \mathbb{1}_{N_h}, \\ \partial_{\mu^{(2)}}\partial_{u^+}\mathcal{L} &= \partial_{\mu^{(3)}}\partial_{u^-}\mathcal{L} = -\mathbb{1}_{N_h\times N_h}, \\ \partial^2_{u^-}\mathcal{L} &= \partial^2_{u^-}J, \\ \partial_{u^+}N^{(1)} = \partial_{u^-}N^{(1)} &= \begin{cases} \kappa\mathbb{1}^\top_{N_h}, & \mu^{(1)} + \kappa(\sum_{i=1}^{N_h} u_i^+ + u_i^- - \alpha) \geq 0, \\ 0, & \text{otherwise}, \end{cases} \\ \partial_{\mu^{(1)}}N^{(1)} &= \begin{cases} 0, & \mu^{(1)} + \kappa(\sum_{i=1}^{N_h} u_i^+ + u_i^- - \alpha) \geq 0, \\ -1, & \text{otherwise}, \end{cases} \\ \partial_{u_j^+}N_i^{(2)} &= \begin{cases} -\kappa\delta_{ij}, & \mu_i^{(2)} - \kappa u_i^+ \geq 0, \\ 0, & \text{otherwise}, \end{cases} \end{aligned}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$$\partial_{\mu_j^{(2)}} N_i^{(2)} = \begin{cases} 0, & \mu_i^{(2)} - \kappa u_i^+ \geq 0, \\ -\delta_{ij}, & \text{otherwise}, \end{cases}$$

$$\partial_{u_j^-} N_i^{(3)} = \begin{cases} -\kappa\delta_{ij}, & \mu_i^{(3)} - \kappa u_i^- \geq 0, \\ 0, & \text{otherwise}, \end{cases}$$

$$\partial_{\mu_j^{(3)}} N_i^{(3)} = \begin{cases} 0, & \mu_i^{(3)} - \kappa u_i^- \geq 0, \\ -\delta_{ij}, & \text{otherwise}. \end{cases}$$

**Numerical example**

Let $\Omega = [0,1]$, $T = 1$, and $a = \frac{1}{100}$. We work on a $20 \times 20$ grid for this example. The positive parts of the measure are displayed by black circles, and the negative parts by red diamonds.

We always start the algorithm with the control being identically zero and terminate when the residual is below $10^{-15}$.

The first example is like that described in Section 6.4.1; compare Figure 6.1. We found the following values to be suitable: the penalty parameter $\gamma = 70$ in $(P_{h,\gamma})$ and the multiplier $\kappa = 2$ to reformulate the KKT-conditions.

The first case we investigate is $\alpha = 0.1$ (see Figure 6.5, top). This $\alpha$ is smaller than the total variation of the true control, and we observe that $\bar{u}^+(\bar{\Omega}) = \alpha$ and $\bar{u}^-(\bar{\Omega}) = 0$. The second case we investigate is $\alpha = 1$ (see Figure 6.5, bottom). This $\alpha$ is equal to the total variation of the true control, and we observe that $\bar{u}^+(\bar{\Omega}) = \alpha$ and $\bar{u}^-(\bar{\Omega}) = 1.8635 \cdot 10^{-20}$. These results are almost identical to those in Section 6.4.1, where only positive measures were allowed (compare Figures 6.2 and 6.3).

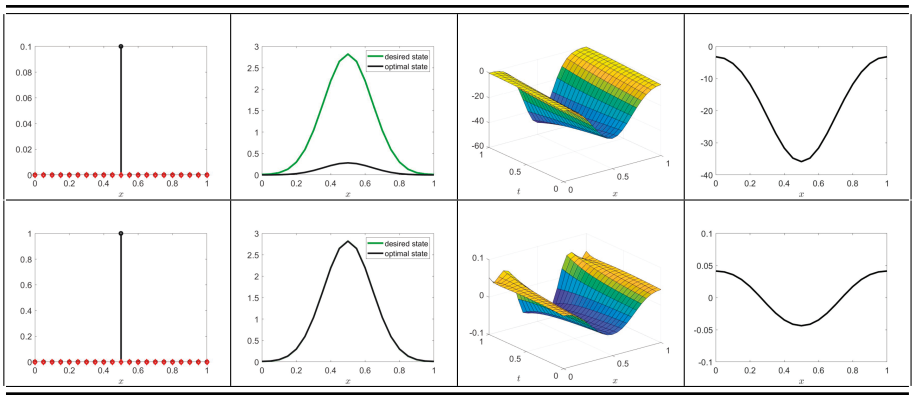



**Figure 6.5:** Solutions for $\alpha = 0.1$ (top) and $\alpha = 1$ (bottom): from left to right: the optimal control $\bar{u} = \bar{u}^+ - \bar{u}^-$ (solved with the semismooth Newton method), associated optimal state $\bar{y}$, associated adjoint $\bar{\varphi}$ on the whole space-time domain $Q$, and associated adjoint $\bar{\varphi}$ at $t = 0$. Terminated after 11 and 64 Newton steps, respectively.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

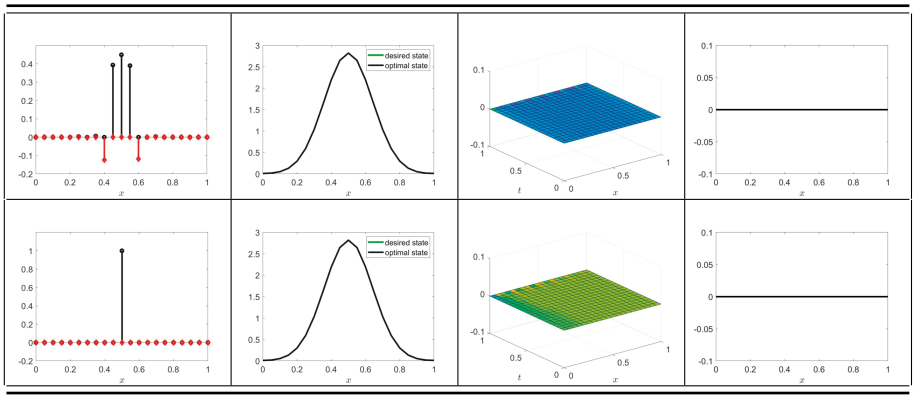



**Figure 6.6:** Solutions for $\alpha = 2$ with original desired state (top) and reachable desired state (bottom): from left to right: the optimal control $\bar{u} = \bar{u}^+ - \bar{u}^-$ (solved with the semismooth Newton method), associated optimal state $\bar{y}$, associated adjoint $\bar{\varphi}$ on the whole space-time domain $Q$, and associated adjoint $\bar{\varphi}$ at $t = 0$. Terminated after 183 and 56 Newton steps, respectively.

The third case we investigate is $\alpha = 2$ (see Figure 6.6). This $\alpha$ is larger than the total variation of the true control, and we observe that $\bar{u}^+(\bar{\Omega}) = 1.5$ and $\bar{u}^-(\bar{\Omega}) = 0.5$. Furthermore, $\bar{y}(T) \approx y_d$ (with an error of size $10^{-8}$), and $\bar{\varphi} \approx 0 \in Q$. Since we allow positive and negative coefficients, the desired state can be reached on the coarse grid – differently from the case of only positive sources, but as a payoff, the sparsity of the optimal control is lost. As required, the complementarity condition has been fulfilled, i. e., $u_i^+ u_i^- = 0$ for all $i$. This, however, comes at the cost of many iterations, since a large constant $\gamma$ causes bad condition of our problem. As a remedy, we implemented a $\gamma$-homotopy like, e. g., in [4, Section 6], where we start with $\gamma = 1$, solve the problem using the semismooth Newton method, and use this solution as a starting point for an increased $\gamma$ until a solution satisfies the constraints. With fixed $\gamma = 70$, we need almost 1000 Newton steps with the $\gamma$-homotopy, which terminates at $\gamma = 64$ in this setting, and it takes 183 Newton steps.

As a comparison to the problem with only positive sources, we also solve the problem with the same reachable desired state as in Figure 6.4, i. e., the projection of the original desired state onto the coarse grid. Here we also observe that $\bar{y}(T) \approx y_d$ (with an error of size $10^{-12}$) and $\bar{\varphi} \approx 0 \in Q$. Furthermore, the optimal control is sparse with $\operatorname{supp}(\bar{u}^+) = \{0.5\}$, it only consists of a positive part, and its total variation is $\bar{u}^+(\bar{\Omega}) = 1 < \alpha$. We fix $\gamma = 70$ and need 56 Newton steps in this case.

Furthermore, we solve this case on a finer mesh ($40 \times 40$) to compare the behavior of solutions (see Figure 6.7). We observe a higher iteration count: 255 Newton steps when employing a $\gamma$-homotopy, which terminates at $\gamma = 64$. In fact, for any example, which we solved on two different meshes, the solver needed more iterations on the finer grid. This is caused by the growing condition number of the PDE solver, since it is a mapping from an initial measure control to the state at the final time. We can also

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

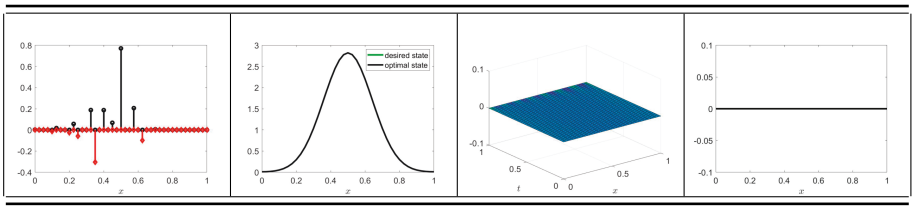



**Figure 6.7:** Solution for $\alpha = 2$ with original desired state on a $40 \times 40$ grid: from left to right: the optimal control $\bar{u} = \bar{u}^+ - \bar{u}^-$ (solved with the semismooth Newton method), associated optimal state $\bar{y}$, associated adjoint $\bar{\varphi}$ on the whole space-time domain $Q$, and associated adjoint $\bar{\varphi}$ at $t = 0$. Terminated after 255 Newton steps.

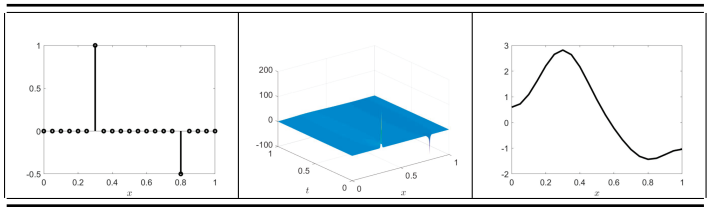


**Figure 6.8:** From left to right: the true solution $u_{\text{true}}$, associated true state $y_{\text{true}}$ in $Q = [0,1] \times [0,1]$, and desired state $y_d = y_{\text{true}}(T)$.

see a difference in the optimal controls in Figure 6.6, top, and Figure 6.7, although a comparable associated optimal state and adjoint are achieved.

The second example we want to look at is a measure consisting of positive and negative parts. To generate the desired state $y_d$, we choose $u_{\text{true}} = \delta_{0.3} - 0.5 \cdot \delta_{0.8}$ and $f \equiv 0$, solve the state equation on a very fine grid ($1000 \times 1000$), and take the evaluation of the result in $t = T$ on the current grid $\Omega_h$ as the desired state $y_d$ (see Figure 6.8).

The first case we investigate is $\alpha = 0.15$ (see Figure 6.9, top). This $\alpha$ is smaller than the total variation of the true control, and we observe that $\bar{u}^+(\bar{\Omega}) = 0.15$ and $\bar{u}^-(\bar{\Omega}) = 1.2929 \cdot 10^{-16}$. The second case we investigate is $\alpha = 1.5$ (see Figure 6.9, bottom). This $\alpha$ is equal to the total variation of the true control, and we observe that $\bar{u}^+(\bar{\Omega}) = 1.0001$ and $\bar{u}^-(\bar{\Omega}) = 0.4999$. For both cases displayed in Figure 6.8, we fix $\gamma = 70$.

Again, we investigate as the third case a setting, where $\alpha = 3 > 1.5 = \|u_{\text{true}}\|_{\mathcal{M}(\bar{\Omega})}$ (see Figure 6.10). We observe that $\bar{u}^+(\bar{\Omega}) = 1.75$ and $\bar{u}^-(\bar{\Omega}) = 1.25$. Here $\bar{y}(T) \approx y_d$ (with an error of size $10^{-7}$), and $\bar{\varphi} \approx 0 \in Q$. The optimal control fulfills the complementarity condition, but we cannot observe the same sparsity inherited by $u_{\text{true}}$. For this case, we had to raise the fixed $\gamma$ to 100, and the computation took over 1700 Newton steps. Hence we employ a $\gamma$-homotopy again, which terminates at $\gamma = 64$ in this setting and only needs 137 Newton steps.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

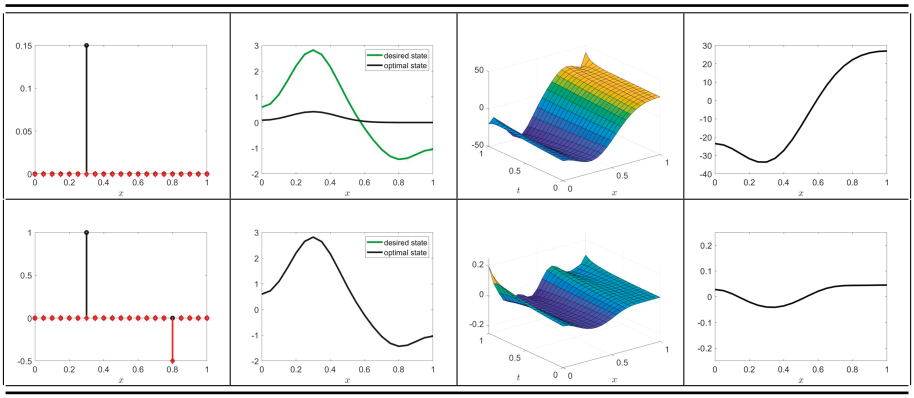



**Figure 6.9:** Solutions for $\alpha = 0.15$ (top) and $\alpha = 1.5$ (bottom): from left to right: the optimal control $\bar{u} = \bar{u}^+ - \bar{u}^-$ (solved with the semismooth Newton method), associated optimal state $\bar{y}$, associated adjoint $\bar{\varphi}$ on the whole space-time domain $Q$, and associated adjoint $\bar{\varphi}$ at $t = 0$. Terminated after 29 and 44 Newton steps, respectively.

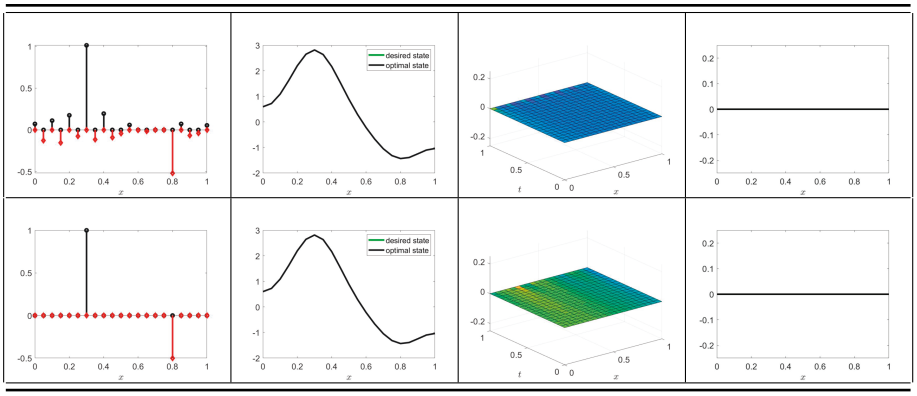



**Figure 6.10:** Solutions for $\alpha = 3$ with original desired state (top) and reachable desired state (bottom): from left to right: the optimal control $\bar{u} = \bar{u}^+ - \bar{u}^-$ (solved with the semismooth Newton method), associated optimal state $\bar{y}$, associated adjoint $\bar{\varphi}$ on the whole space-time domain $Q$, and associated adjoint $\bar{\varphi}$ at $t = 0$. Terminated after 137 and 20 Newton steps.

For comparison, we project the desired state onto the coarse grid so that it becomes reachable and then solve the problem again. Now we observe that $\bar{u}^+(\bar{\Omega}) = 1$, $\bar{u}^-(\bar{\Omega}) = 0.5$, $\mathrm{supp}(\bar{u}^+) = \{0.3\}$, and $\mathrm{supp}(\bar{u}^-) = \{0.8\}$, which are exactly the properties of $u_{\mathrm{true}}$. Furthermore, we see that $\bar{y}(T) \approx y_d$ (with an error of size $10^{-14}$) and $\bar{\varphi} \approx 0 \in Q$. We observe a reduction of Newton steps needed: the computation took 20 Newton steps with fixed $\gamma = 100$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

# Bibliography


[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[2] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer, New York, 2011.

[3] E. Casas, C. Clason, and K. Kunisch. Approximation of elliptic control problems in measure spaces with sparse solutions. *SIAM J. Control Optim.*, 50(4):1735–1752, 2012.

[4] E. Casas, C. Clason, and K. Kunisch. Parabolic control problems in measure spaces with sparse solutions. *SIAM J. Control Optim.*, 51(1):28–63, 2013.

[5] E. Casas and K. Kunisch. Parabolic control problems in space-time measure spaces. *ESAIM Control Optim. Calc. Var.*, 22(2):355–370, 2016.

[6] E. Casas and K. Kunisch. Using sparse control methods to identify sources in linear diffusion–convection equations. *Inverse Probl.*, 35(11):114002, 2019.

[7] E. Casas, B. Vexler, and E. Zuazua. Sparse initial data identification for parabolic PDE and its finite element approximations. *Math. Control Relat. Fields*, 5(3):377–399, 2015.

[8] C. Clason and K. Kunisch. A duality-based approach to elliptic control problems in non-reflexive Banach spaces. *ESAIM Control Optim. Calc. Var.*, 17(1):243–266, 2011.

[9] C. Clason and A. Schiela. Optimal control of elliptic equations with positive measures. *ESAIM Control Optim. Calc. Var.*, 23(1):217–240, 2017.

[10] A. El Badia, T. Ha-Duong, and A. Hamdi. Identification of a point source in a linear advection–dispersion–reaction equation: application to a pollution source problem. *Inverse Probl.*, 21(3):1121, 2005.

[11] W. Gong. Error estimates for finite element approximations of parabolic equations with measure data. *Math. Comput.*, 82(281):69–98, 2013.

[12] W. Gong, M. Hinze, and Z. Zhou. A priori error analysis for finite element approximation of parabolic optimal control problems with pointwise control. *SIAM J. Control Optim.*, 52(1):97–119, 2014.

[13] S.-P. Han and O. L. Mangasarian. Exact penalty functions in nonlinear programming. *Math. Program.*, 17(1):251–269, 1979.

[14] E. Herberg, M. Hinze, and H. Schumacher. Maximal discrete sparsity in parabolic optimal control with measures. *arXiv preprint* arXiv:1804.10549, 2018.

[15] M. Hinze. A variational discretization concept in control constrained optimization: the linear-quadratic case. *Comput. Optim. Appl.*, 30(1):45–61, 2005.

[16] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints*, volume 23 of Mathematical Modelling: Theory and Applications. Springer, New York, 2009.

[17] K. Kunisch, K. Pieper, and B. Vexler. Measure valued directional sparsity for parabolic optimal control problems. *SIAM J. Control Optim.*, 52(5):3078–3108, 2014.

[18] D. Leykekhman, B. Vexler, and D. Walter. Numerical analysis of sparse initial data identification for parabolic problems. *arXiv preprint* arXiv:1905.01226, 2019.

[19] Y. Li, S. Osher, and R. Tsai. Heat source identification based on constrained minimization. *Inverse Probl. Imaging*, 8(1):199–221, 2014.

[20] K. Pieper and B. Vexler. A priori error analysis for discretization of sparse elliptic optimal control problems in measure space. *SIAM J. Control Optim.*, 51(4):2788–2808, 2013.

[21] G. Stadler. Elliptic optimal control problems with $L^1$-control cost and applications for the placement of control devices. *Comput. Optim. Appl.*, 44(2):159–181, 2009.


EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Adrian Hirn and Winnifried Wollner

# 7 An optimal control problem for equations with *p*-structure and its finite element discretization

**Abstract:** We analyze a finite element approximation of an optimal control problem that involves an elliptic equation with $p$-structure (e. g., the $p$-Laplace) as a constraint. As the nonlinear operator related to the $p$-Laplace equation mapping the space $W_0^{1,p}(\Omega)$ to its dual $(W_0^{1,p}(\Omega))^*$ is not Gâteaux differentiable, first-order optimality conditions cannot be formulated in a standard way. Without using adjoint information, we derive novel a priori error estimates for the convergence of the cost functional for both variational discretization and piecewise constant controls.

**Keywords:** $p$-Laplacian, optimization, finite element method, optimality conditions, a priori error estimates

**MSC 2010:** 49K20, 49M25, 65N30, 65N12

## 7.1 Introduction

In this paper, for given $\alpha > 0$ and $u_d \in L^2(\Omega)$, we study the finite element discretization of the following elliptic optimal control problem:

$$\text{Minimize} \quad J(q,u) := \frac{1}{2}\|u - u_d\|_2^2 + \frac{\alpha}{2}\|q\|_2^2 \tag{7.1a}$$

subject to the PDE-constraints

$$-\operatorname{div} \boldsymbol{S}(\nabla u) = q \quad \text{in } \Omega, \tag{7.1b}$$

$$u = 0 \quad \text{on } \partial\Omega, \tag{7.1c}$$

**Acknowledgement:** The authors would like to thank the anonymous referees for their valuable comments helping to shorten and improve the manuscript.
W. Wollner is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 314067056.

**Adrian Hirn,** Hochschule Esslingen, Robert-Bosch-Straße 1, 73037 Göppingen, Germany, e-mail: adrian.hirn@hs-esslingen.de
**Winnifried Wollner,** Technische Universität Darmstadt, Fachbereich Mathematik, Dolivostr. 15, 64293 Darmstadt, Germany, e-mail: wollner@mathematik.tu-darmstadt.de

https://doi.org/10.1515/9783110695984-007

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

and for $q_a, q_b \in \mathbb{R}$, $q_a < q_b$, and, without loss of generality, $0 \in [q_a, q_b]$, the box-constraints

$$q_a \leq q(x) \leq q_b \quad \text{for a. a. } x \in \Omega, \tag{7.1d}$$

where for given $p > 1$ and $\varepsilon \geq 0$, the nonlinear vector field $\boldsymbol{S} : \mathbb{R}^d \to \mathbb{R}^d$ is supposed to have $p$-structure or, more precisely, $(p, \varepsilon)$-structure; see Assumption 7.2.1. Prototypical examples falling into this class are

$$\boldsymbol{S}(\nabla u) = |\nabla u|^{p-2}\nabla u \quad \text{or} \quad \boldsymbol{S}(\nabla u) = \left(\varepsilon^2 + |\nabla u|^2\right)^{\frac{p-2}{2}} \nabla u. \tag{7.2}$$

Throughout the paper, $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, is either a convex polyhedral domain or a bounded convex domain with smooth boundary $\partial\Omega \in C^2$. We include the case of curved boundary in our analysis as we need to assume certain conditions on the regularity of $u$ that are so far only available for domains with smooth boundaries. For simplicity of exposition, we restrict the analysis to $d = 2$ in case of a curved boundary.

Equations with $p$-structure arise in various physical applications, such as the theory of plasticity, bimaterial problems in elastic-plastic mechanics, non-Newtonian fluid mechanics, blood rheology, and glaciology; see, e. g., [26, 32, 34] and the references therein. The first operator in (7.2) corresponds to the $p$-Laplace equation. For $\varepsilon > 0$, the second operator in (7.2) regularizes the degeneracy of the $p$-Laplacian as the modulus of the gradient tends to zero. For $\varepsilon = 0$, $(7.2)_2$ reduces to the $p$-Laplace operator $(7.2)_1$. Finite element (FE) approximations of the $p$-Laplace equation and related equations have been widely investigated; see [1, 22, 24, 30]. Attention to optimal control of quasi-linear PDEs is given, e. g., in [9, 11]. The extension to the parabolic case can be found in [6, 8]. The approximation of optimal control problems in the coefficient for the $p$-Laplacian by its $\varepsilon$-regularization is studied in [10]. For problem (7.1), some DWR-type a posteriori error estimates have been utilized in [25]. To the best knowledge of the authors, no a priori discretization error results are available for this problem class; by our work we want to fill this gap. We point out that the case $\varepsilon = 0$ is included in our analysis.

A standard procedure for the finite element analysis of an optimal control problem consists in deriving first-order optimality conditions and exploiting the properties of the adjoint state. Unfortunately, this analysis requires the existence of a suitable adjoint state. To see that this cannot be guaranteed by standard theory, note that the nonlinear operator related to the $p$-Laplace equation maps the space $W_0^{1,p}(\Omega)$ to its dual $(W_0^{1,p}(\Omega))^*$. Hence its formal derivative is a linear operator mapping $W_0^{1,p}(\Omega)$ into its dual. As it can be seen in the calculations yielding (7.36), the corresponding linear operator is positive and thus injective; further, it is clearly self-adjoint; see (7.31). Hence, unless $p = 2$, the linear operator cannot be surjective; see, e. g., the discussion in [27]. Hence standard KKT-theory is not applicable in the natural setting.

Despite this lack of standard theory, for $\varepsilon > 0$, we are able to show the existence of a suitable discrete adjoint state allowing a discrete optimality system suitable for

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

a variational discretization in the spirit of [28]. Due to lack of first-order optimality conditions on the continuous level, we cannot attain additional regularity of the adjoint variable. Without additional regularity of these variables, we cannot expect more than qualitative convergence for them. Hence, to establish a priori error estimates, we follow techniques established for elliptic optimization problems with state [18, 35] or gradient-state [36] constraints, where also no convergence rates of the adjoint variable are available. Although in our analysis, we can adopt ideas from [36], we have to cope with several challenges due to the nonlinear degenerate PDE-constraint. For discretization of (7.1), we consider two possible approaches: (a) variational discretization with piecewise linear states and (b) piecewise linear states and piecewise constant controls. In case of (a) the control space is discretized implicitly by the discrete adjoint equation. We show that the sequence of discrete global minimizers $(\overline{q}_h, \overline{u}_h)$ for mesh size $h \in (0, 1]$ has a strong accumulation point $(\overline{q}, \overline{u})$ that is a global optimal solution to (7.1). Under a certain realistic regularity assumption for solutions of the state equation (Assumption 7.2.4), we prove a quantitative convergence estimate for the cost functional value for both variational discretization and piecewise constant controls in Theorems 7.7.2 and 7.7.3.

For the proof of these estimates, we combine methods from [36] with quasi-norm techniques from [22] to handle the degeneracy of the nonlinear operator. Our method does not require additional regularity of the control variable. The required regularity in Assumption 7.2.4 is verified for the $p$-Laplace equation on bounded convex domains with $C^2$-boundary in [13, 16].

The plan of the paper is as follows. In Section 7.2, we fix our notation and clarify the structure of the nonlinear vector field $\boldsymbol{S}$. Further, we state our assumption on the regularity of solutions to (7.1b)–(7.1c) (Assumption 7.2.4). Section 7.3 is concerned with the precise formulation of the optimal control problem (7.1) and its solvability. In Section 7.4, we describe its finite element discretization, followed in Section 7.5 by an analysis of the first-order optimality conditions. In Section 7.6, we collect and extend several results on the finite element approximation of the $p$-Laplace equation to apply them in Section 7.7 to the convergence analysis of the optimal control problem. There we verify without any regularity assumption that the sequence of discrete minimizers $(\overline{q}_h, \overline{u}_h)$ has a strong accumulation point $(\overline{q}, \overline{u})$, which is an optimal solution to (7.1). Under the regularity Assumption 7.2.4, we then prove a priori error estimates quantifying the order of convergence in the cost functional.

## 7.2 Preliminaries

To begin with, we clarify our notation and state important properties of the nonlinear operator in (7.1b). Further, we pose our assumption on the regularity of solutions to the state equation (7.1b)–(7.1c), which will be crucial for our analysis.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

### 7.2.1 Notation

The set of all positive real numbers is denoted by $\mathbb{R}^+$. Let $\mathbb{R}_0^+ := \mathbb{R}^+ \cup \{0\}$. The Euclidean scalar product of two vectors $\boldsymbol{\xi}, \boldsymbol{\eta} \in \mathbb{R}^d$ is denoted by $\boldsymbol{\xi} \cdot \boldsymbol{\eta}$. We set $|\boldsymbol{\eta}| := (\boldsymbol{\eta} \cdot \boldsymbol{\eta})^{1/2}$. We often use $c$ as a generic constant whose value may change from line to line but does not depend on important variables. We write $a \sim b$ if there exist constants $c, C > 0$, independent of all relevant quantities, such that $cb \le a \le Cb$. Similarly, the notation $a \lesssim b$ stands for $a \le Cb$.

Let $\omega \subset \Omega$ be a measurable nonempty set. The $d$-dimensional Lebesgue measure of $\omega$ is denoted by $|\omega|$. The mean value of a Lebesgue-integrable function $f$ over $\omega$ is denoted by

$$\langle f \rangle_\omega := \fint_\omega f(x)\,\mathrm{d}x := \frac{1}{|\omega|}\int_\omega f(x)\,\mathrm{d}x.$$

For $\nu \in [1,\infty]$, $L^\nu(\Omega)$ stands for the Lebesgue space, and $W^{m,\nu}(\Omega)$ for the Sobolev space of order $m$. For $\nu > 1$, we denote by $W_0^{1,\nu}(\Omega)$ the Sobolev space with vanishing traces on $\partial\Omega$. The $L^\nu(\omega)$-norm is denoted by $\|\cdot\|_{\nu;\omega}$, and the $W^{m,\nu}(\omega)$-norm is denoted by $\|\cdot\|_{m,\nu;\omega}$. For $\nu \in (1,\infty)$ and $\frac{1}{\nu} + \frac{1}{\nu'} = 1$, i. e., $\nu' = \frac{\nu}{\nu-1}$, the dual space of $W_0^{1,\nu}(\omega)$ is denoted by $W^{-1,\nu'}(\omega) = (W_0^{1,\nu}(\omega))^*$, and for its dual norm, we write $\|\cdot\|_{-1,\nu';\omega}$. For the $L^2(\omega)$ inner product, we use the notation $(\cdot,\cdot)_\omega$. This notation of norms and inner products is also used for vector-valued functions. In case of $\omega = \Omega$, we usually omit the index $\Omega$, e. g., $\|\cdot\|_\nu = \|\cdot\|_{\nu;\Omega}$.

We recall the important Poincaré inequality: For $\nu \in (1,\infty)$, we have

$$\|u\|_{\nu;\omega} \le c_P \|\nabla u\|_{\nu;\omega} \quad \forall u \in W_0^{1,\nu}(\omega). \tag{7.3}$$

There exist diverse generalizations of Poincaré's inequality. We will make use of the following version, which goes back to [5]: Let $\omega$ be a bounded convex open subset of $\mathbb{R}^d$, $d \ge 1$, and let $\varphi : [0,\infty) \to [0,\infty)$ be continuous and convex with $\varphi(0) = 0$. Let $u : \omega \to \mathbb{R}^N$, $N \ge 1$, be in $W^{1,1}(\omega)$ such that $\varphi(|\nabla u|) \in L^1(\omega)$. Then

$$\int_\omega \varphi\left(\frac{|u(x) - \langle u\rangle_\omega|}{\delta}\right)\mathrm{d}x \le \left(\frac{V_d \delta^d}{|\omega|}\right)^{1-\frac{1}{d}} \int_\omega \varphi\big(|\nabla u(x)|\big)\,\mathrm{d}x, \tag{7.4}$$

where $\delta$ is the diameter of $\omega$, and $V_d$ is the volume of the unit ball in $\mathbb{R}^d$.

### 7.2.2 Properties of the nonlinear operator

In this section, we state our assumptions on the nonlinear operator $\boldsymbol{S}$. Further, we discuss important properties of the nonlinear operator, and we indicate how it relates to so-called N-functions.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Assumption 7.2.1** (Nonlinear operator). We assume that the nonlinear operator $\boldsymbol{S} : \mathbb{R}^d \to \mathbb{R}^d$ belongs to $C^0(\mathbb{R}^d, \mathbb{R}^d) \cap C^1(\mathbb{R}^d \setminus \{\mathbf{0}\}, \mathbb{R}^d)$ and satisfies $\boldsymbol{S}(\mathbf{0}) = \mathbf{0}$. Furthermore, we assume that the operator $\boldsymbol{S}$ possesses $(p, \varepsilon)$-structure, i. e., there exist $p \in (1, \infty)$, $\varepsilon \in [0, \infty)$, and constants $C_0, C_1 > 0$ such that

$$\sum_{i,j=1}^{d} \partial_i S_j(\boldsymbol{\xi}) \eta_i \eta_j \geq C_0 (\varepsilon + |\boldsymbol{\xi}|)^{p-2} |\boldsymbol{\eta}|^2, \tag{7.5a}$$

$$\left|\partial_i S_j(\boldsymbol{\xi})\right| \leq C_1 (\varepsilon + |\boldsymbol{\xi}|)^{p-2} \tag{7.5b}$$

for all $\boldsymbol{\xi}, \boldsymbol{\eta} \in \mathbb{R}^d$ with $\boldsymbol{\xi} \neq \mathbf{0}$ and all $i, j \in \{1, \dots, d\}$.

Important examples of nonlinear operators $\boldsymbol{S}$ satisfying Assumption 7.2.1 are those derived from a potential with $(p, \varepsilon)$-structure, i. e., there exists a convex function $\Phi : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ belonging to $C^1(\mathbb{R}_0^+) \cap C^2(\mathbb{R}^+)$ and satisfying $\Phi(0) = 0$ and $\Phi'(0) = 0$ such that for all $\boldsymbol{\xi} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ and $i = 1, \dots, d$,

$$S_i(\boldsymbol{\xi}) = \partial_i(\Phi(|\boldsymbol{\xi}|)) = \Phi'(|\boldsymbol{\xi}|) \frac{\xi_i}{|\boldsymbol{\xi}|}. \tag{7.6}$$

If, in addition, $\Phi$ possesses $(p, \varepsilon)$-structure, i. e., if there exist $p \in (1, \infty)$, $\varepsilon \in [0, \infty)$, and constants $C_2, C_3 > 0$ such that for all $t > 0$,

$$C_2 (\varepsilon + t)^{p-2} \leq \Phi''(t) \leq C_3 (\varepsilon + t)^{p-2}, \tag{7.7}$$

then we can show (see [4]) that $\boldsymbol{S}$ satisfies Assumption 7.2.1. Note that (7.2) falls into this class. We will briefly discuss how the operator $\boldsymbol{S}$ with $(p, \varepsilon)$-structure relates to N-functions that are standard in the theory of Orlicz spaces; see [20]. We define the convex function $\varphi : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ by

$$\varphi(t) := \int_0^t (\varepsilon + s)^{p-2} s \, \mathrm{d}s. \tag{7.8}$$

The function $\varphi$ belongs to $C^1(\mathbb{R}_0^+) \cap C^2(\mathbb{R}^+)$ and satisfies, uniformly in $t > 0$,

$$\min\{1, p-1\} (\varepsilon + t)^{p-2} \leq \varphi''(t) \leq \max\{1, p-1\} (\varepsilon + t)^{p-2}. \tag{7.9}$$

Therefore inequalities (7.5a) and (7.5b) defining the $(p, \varepsilon)$-structure of $\boldsymbol{S}$ can be expressed equivalently in terms of the convex function $\varphi$:

$$\sum_{i,j=1}^{d} \partial_i S_j(\boldsymbol{\xi}) \eta_i \eta_j \geq \widetilde{C}_0 \varphi''(|\boldsymbol{\xi}|) |\boldsymbol{\eta}|^2, \quad \left|\partial_i S_j(\boldsymbol{\xi})\right| \leq \widetilde{C}_1 \varphi''(|\boldsymbol{\xi}|) \quad \forall i, j = 1, \dots, d.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

The function $\varphi$ is an example of an N-function satisfying the $\Delta_2$-condition; see, e. g., [20, 23]. In view of (7.9), the function $\varphi$ satisfies uniformly in $t$ the equivalence

$$\varphi''(t)t \sim \varphi'(t). \tag{7.10}$$

Several studies on the finite element analysis of the $p$-Laplace equation indicate that a $p$-structure-adapted quasi-norm is crucial for error estimation. To this end, for given $\psi \in C^1([0,\infty))$, we introduce the family of shifted functions $\{\psi_a\}_{a\geq 0}$ by

$$\psi_a(t) := \int_0^t \psi_a'(s)\,\mathrm{d}s \quad \text{with} \quad \psi_a'(t) := \frac{\psi'(a+t)}{a+t}t. \tag{7.11}$$

For $\psi = \varphi$ given by (7.8), we have $\varphi_a(t) \sim (\varepsilon + a + t)^{p-2}t^2$ uniformly in $t \geq 0$. In [20] the following Young-type inequality is provided: For all $\delta > 0$, there exists $c(\delta) > 0$ such that for all $s$, $t$, $a \geq 0$,

$$s\varphi_a'(t) + \varphi_a'(s)t \leq \delta\varphi_a(s) + c(\delta)\varphi_a(t), \tag{7.12}$$

where the constant $c(\delta)$ only depends on $p$ and $\delta$ (it is independent of $\varepsilon$ and $a$).

We define the function $\boldsymbol{F} : \mathbb{R}^d \to \mathbb{R}^d$ associated with the nonlinear operator $\boldsymbol{S}$ with $(p,\varepsilon)$-structure by

$$\boldsymbol{F}(\boldsymbol{\xi}) := \left(\varepsilon + |\boldsymbol{\xi}|\right)^{\frac{p-2}{2}}\boldsymbol{\xi}, \tag{7.13}$$

where $p$ and $\varepsilon$ are the same as in Assumption 7.2.1. The vector fields $\boldsymbol{S}$ and $\boldsymbol{F}$ are closely related to each other as depicted by the following lemma provided by [19, 20].

**Lemma 7.2.2.** *For $p \in (1,\infty)$ and $\varepsilon \in [0,\infty)$, let $\boldsymbol{S}$ satisfy Assumption* 7.2.1*, and let $\boldsymbol{F}$, $\varphi$, and $\varphi_{|\boldsymbol{\xi}|}$ be defined by* (7.13)*,* (7.8)*, and* (7.11)*, respectively. Then for all $\boldsymbol{\xi}, \boldsymbol{\eta} \in \mathbb{R}^d$, we have*

$$\begin{aligned}(\boldsymbol{S}(\boldsymbol{\xi}) - \boldsymbol{S}(\boldsymbol{\eta}))\cdot(\boldsymbol{\xi}-\boldsymbol{\eta}) &\sim \left(\varepsilon + |\boldsymbol{\xi}| + |\boldsymbol{\eta}|\right)^{p-2}|\boldsymbol{\xi}-\boldsymbol{\eta}|^2\\ &\sim \varphi_{|\boldsymbol{\xi}|}(|\boldsymbol{\xi}-\boldsymbol{\eta}|) \sim \left|\boldsymbol{F}(\boldsymbol{\xi}) - \boldsymbol{F}(\boldsymbol{\eta})\right|^2,\\ \left|\boldsymbol{S}(\boldsymbol{\xi}) - \boldsymbol{S}(\boldsymbol{\eta})\right| &\sim \varphi_{|\boldsymbol{\xi}|}'(|\boldsymbol{\xi}-\boldsymbol{\eta}|) \sim \left(\varepsilon + |\boldsymbol{\xi}| + |\boldsymbol{\eta}|\right)^{p-2}|\boldsymbol{\xi}-\boldsymbol{\eta}|,\end{aligned}$$

*where the constants only depend on $p$; in particular, they are independent of $\varepsilon \geq 0$.*

Due to Lemma 7.2.2, for all $u, v \in W^{1,p}(\Omega)$, we have the equivalence

$$\left(\boldsymbol{S}(\nabla u) - \boldsymbol{S}(\nabla v), \nabla u - \nabla v\right)_\Omega \sim \left\|\boldsymbol{F}(\nabla u) - \boldsymbol{F}(\nabla v)\right\|_2^2 \sim \int_\Omega \varphi_{|\nabla u|}(|\nabla u - \nabla v|)\,\mathrm{d}x \tag{7.14}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

with constants only depending on $p$. We refer to each quantity in (7.14) as the *quasi-norm* or *natural distance* following, e. g., [1–3, 21]. It has been used very successfully in the finite element analysis of equations with $p$-structure.

The following lemma from [29] shows the connection between the natural distance and the Sobolev norms.

**Lemma 7.2.3.** *For $p \in (1,\infty)$ and $\varepsilon \in [0,\infty)$, let the operator $\boldsymbol{S}$ satisfy Assumption 7.2.1, and let $\boldsymbol{F}$ be defined by (7.13). Then for all $u, v \in W^{1,p}(\Omega)$, we have:*
(i) *in the case $p \in (1,2]$, with constants only depending on $p$,*

$$\begin{aligned}\left\|\nabla(u-v)\right\|_p^2 &\lesssim \left\|\boldsymbol{F}(\nabla u)-\boldsymbol{F}(\nabla v)\right\|_2^2 \left\|\varepsilon+|\nabla u|+|\nabla v|\right\|_p^{2-p},\\ \left\|\boldsymbol{F}(\nabla u)-\boldsymbol{F}(\nabla v)\right\|_2^2 &\lesssim \left\|\nabla(u-v)\right\|_p^p;\end{aligned}$$

(ii) *in the case $p \in [2,\infty)$, with constants only depending on $p$,*

$$\left\|\nabla(u-v)\right\|_p^p \lesssim \left\|\boldsymbol{F}(\nabla u)-\boldsymbol{F}(\nabla v)\right\|_2^2 \lesssim \left\|\varepsilon+|\nabla u|+|\nabla v|\right\|_p^{p-2} \left\|\nabla(u-v)\right\|_p^2.$$

*In particular, all constants appearing in* (i) *and* (ii) *are independent of $\varepsilon \geq 0$.*

## 7.2.3 Regularity assumption

We impose our assumption on the regularity of solutions to the state equation, which will later enable us to derive a priori error estimates for the finite element approximation of (7.1).

**Assumption 7.2.4.** Let $q \in L^{\max\{2,p'\}}(\Omega)$, $\frac{1}{p}+\frac{1}{p'}=1$. Then the weak solution $u$ to equation (7.1b)–(7.1c) with $p$-structure satisfies the regularity

$$\boldsymbol{S}(\nabla u) \in W^{1,2}(\Omega) \quad \text{and} \quad u \in W^{2,2}(\Omega),$$

and there exist positive constants $c_1$, $c_2$, $\gamma$ such that

$$\left\|\boldsymbol{S}(\nabla u)\right\|_{1,2} \leq c_1 \|q\|_2 \quad \text{and} \quad \|u\|_{2,2} \leq c_2 \|q\|_{\max\{2,p'\}}^{\gamma}. \tag{7.15}$$

The regularity Assumption 7.2.4 is satisfied for certain data:

– In [13], it is shown that if $\Omega \subset \mathbb{R}^d$, $d \geq 2$, is a bounded convex open set, $q \in L^2(\Omega)$, and $p \in (1,\infty)$, then the weak solution $u$ to equation (7.1b)–(7.1c) with $p$-structure satisfies the regularity $\boldsymbol{S}(\nabla u) \in W^{1,2}(\Omega)$ with

$$C_1\|q\|_2 \leq \left\|\boldsymbol{S}(\nabla u)\right\|_{1,2} \leq C_2\|q\|_2,$$

where the constants $C_1$, $C_2$ only depend on $p$ and $d$. In particular, the analysis carried out in [13] covers the $p$-Laplacian with $\boldsymbol{S}(\nabla u) = |\nabla u|^{p-2}\nabla u$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

- In [14], on certain domains, Lipschitz-continuous solutions are obtained whenever $q \in L^r(\Omega)$ with $r > d$ has mean-value zero and $\varepsilon > 0$. [13, Remark 2.7] claims that this implies the $W^{2,2}$-regularity.
- In [16], it is shown that, if $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, is a bounded domain with $C^2$-boundary, $p \in (1,2]$, and $q \in L^{p'}(\Omega)$, then the weak solution $u$ of the $p$-Laplace equation (7.1b)–(7.1c) with $\boldsymbol{S}(\nabla u) = |\nabla u|^{p-2}\nabla u$ fulfills

$$u \in W^{2,2}(\Omega) \quad \text{with} \quad \left\|\nabla^2 u\right\|_2 \leq C\|q\|_{p'}^{\frac{1}{p-1}}.$$

- In [33], it is shown that if $\Omega \subset \mathbb{R}^2$ is either convex or has $C^2$-boundary, then for $p \in (1,2)$, the weak solution $u$ of the $p$-Laplace equation (7.1b)–(7.1c) with $\boldsymbol{S}(\nabla u) = |\nabla u|^{p-2}\nabla u$ satisfies

$$q \in L^r \text{ with } r > 2 \quad \Longrightarrow \quad u \in W^{2,2}(\Omega).$$

As a consequence, Assumption 7.2.4 is satisfied for the $p$-Laplace equation in the case $p \in (1,2]$ if, e. g., $\Omega \subset \mathbb{R}^d$, $d \geq 2$, is a bounded convex domain with $C^2$-boundary and $q \in L^{p'}(\Omega)$. Since later we take $q \in \mathcal{Q}_{\text{ad}} \subset L^\infty(\Omega)$, we can weaken Assumption 7.2.4: It is sufficient to assume that $q \in L^\infty(\Omega)$ implies $\boldsymbol{S}(\nabla u) \in W^{1,2}(\Omega)$ and $u \in W^{2,2}(\Omega)$ with $\|\boldsymbol{S}(\nabla u)\|_{1,2} \lesssim \|q\|_\infty$ and $\|u\|_{2,2} \lesssim \|q\|_\infty^\gamma$ (replacing (7.15)).

## 7.3 Optimal control problem

In this section, we give a precise definition of the optimal control problem (7.1). For $\frac{1}{p} + \frac{1}{p'} = 1$, i. e., $p' = \frac{p}{p-1}$, the natural spaces for the states and controls are

$$\mathcal{V} := W_0^{1,p}(\Omega), \quad Q := L^{\max\{2,p'\}}(\Omega), \quad \mathcal{Q}_{\text{ad}} := \{q \in Q \mid q_a \leq q \leq q_b \text{ a. e. in } \Omega\}.$$

The weak formulation of the state equation (7.1b)–(7.1c) reads as follows:

For a given control $q \in \mathcal{Q}_{\text{ad}}$, find the state $u = u(q) \in \mathcal{V}$ such that

$$\left(\boldsymbol{S}(\nabla u), \nabla \varphi\right)_\Omega = (q, \varphi)_\Omega \quad \forall \varphi \in \mathcal{V}. \tag{7.16}$$

We now investigate stability and continuity properties of the solution $u \in \mathcal{V}$ with respect to the control $q$. It will be suitable for this to consider variations of $q$ in $W^{-1,p'}(\Omega)$.

**Lemma 7.3.1.** *For all $p \in (1,\infty)$ and $q \in \mathcal{Q}_{\text{ad}}$, there exists a unique solution $u = u(q) \in \mathcal{V}$ to* (7.16). *This solution satisfies the a priori estimate*

$$\|\nabla u\|_p \leq c_1\left(\|q\|_{-1,p'}^{\frac{1}{p-1}} + c_2\varepsilon\right), \tag{7.17}$$

*where $c_1 > 0$ only depends on $\Omega$ and $p$, and $c_2 = 1$ if $p < 2$ and $c_2 = 0$ otherwise.*

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*Proof.* Lemma 7.2.2 implies that the operator $-\operatorname{div} \boldsymbol{S}(\nabla\cdot) : W_0^{1,p}(\Omega) \to W^{-1,p'}(\Omega)$ is strictly monotone. Using the theory of monotone operators (see [38, 41]), we can thus easily conclude that for each $q \in Q$, there exists a unique solution $u = u(q) \in \mathcal{V}$ to (7.16). The proof of (7.17) is standard and can be found, e. g., in [29]. □

The next lemma states that the solution operator $q \mapsto u = u(q)$ is locally Hölder-continuous.

**Lemma 7.3.2.** *For $p \in (1,\infty)$, let $u_1 = u(q_1) \in \mathcal{V}$ and $u_2 = u(q_2) \in \mathcal{V}$ be the solutions to the state equation* (7.16) *for the right-hand side $q_1 \in \mathcal{Q}_{\mathrm{ad}}$ and $q_2 \in \mathcal{Q}_{\mathrm{ad}}$. Then there exist constants, only depending on $p$, $\Omega$ such that*

$$\left\|\boldsymbol{F}(\nabla u_1) - \boldsymbol{F}(\nabla u_2)\right\|_2 \lesssim \begin{cases} \|\varepsilon + |\nabla u_1| + |\nabla u_2|\|_p^{\frac{2-p}{2}} \|q_1 - q_2\|_{-1,p'} & \text{for } p \le 2, \\ \|q_1 - q_2\|_{-1,p'}^{\frac{p'}{2}} & \text{for } p \ge 2, \end{cases}$$

$$\|\nabla u_1 - \nabla u_2\|_p \lesssim \begin{cases} \|\varepsilon + |\nabla u_1| + |\nabla u_2|\|_p^{2-p} \|q_1 - q_2\|_{-1,p'} & \text{for } p \le 2, \\ \|q_1 - q_2\|_{-1,p'}^{\frac{1}{p-1}} & \text{for } p \ge 2. \end{cases}$$

*Proof.* As $u_1 \in \mathcal{V}$ and $u_2 \in \mathcal{V}$ solve (7.16) with right-hand sides $q_1$ and $q_2$, we have

$$\left(\boldsymbol{S}(\nabla u_1) - \boldsymbol{S}(\nabla u_2), \nabla\varphi\right)_\Omega = \langle q_1 - q_2, \varphi\rangle \quad \forall \varphi \in W_0^{1,p}(\Omega).$$

Testing this equation with $\varphi = u_1 - u_2$ and employing Lemma 7.2.2, we get

$$\left\|\boldsymbol{F}(\nabla u_1) - \boldsymbol{F}(\nabla u_2)\right\|_2^2 \sim \langle q_1 - q_2, u_1 - u_2\rangle \le \|q_1 - q_2\|_{-1,p'} \|u_1 - u_2\|_{1,p} \tag{7.18}$$

for $p \in (1,\infty)$. Poincaré's inequality and Lemma 7.2.3 imply

$$\begin{aligned} &\left\|\boldsymbol{F}(\nabla u_1) - \boldsymbol{F}(\nabla u_2)\right\|_2^2 \\ &\quad\lesssim \|q_1 - q_2\|_{-1,p'} \|\nabla u_1 - \nabla u_2\|_p \\ &\quad\lesssim \begin{cases} \|q_1 - q_2\|_{-1,p'} \|\varepsilon + |\nabla u_1| + |\nabla u_2|\|_p^{\frac{2-p}{2}} \|\boldsymbol{F}(\nabla u_1) - \boldsymbol{F}(\nabla u_2)\|_2 & \text{for } p \le 2, \\ \|q_1 - q_2\|_{-1,p'} \|\boldsymbol{F}(\nabla u_1) - \boldsymbol{F}(\nabla u_2)\|_2^{\frac{2}{p}} & \text{for } p \ge 2. \end{cases} \end{aligned}$$

This yields the desired estimate in the natural distance. Using similar arguments, we have

$$\begin{aligned} \|q_1 - q_2\|_{-1,p'} \|\nabla u_1 - \nabla u_2\|_p &\overset{(7.18)}{\gtrsim} \left\|\boldsymbol{F}(\nabla u_1) - \boldsymbol{F}(\nabla u_2)\right\|_2^2 \\ &\gtrsim \begin{cases} \|\varepsilon + |\nabla u_1| + |\nabla u_2|\|_p^{p-2} \|\nabla u_1 - \nabla u_2\|_p^2 & \text{for } p \le 2, \\ \|\nabla u_1 - \nabla u_2\|_p^p & \text{for } p \ge 2. \end{cases} \end{aligned}$$

From this we obtain the desired estimate in the $W_0^{1,p}$-norm. □

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

For given $\alpha > 0$ and $u_d \in L^2(\Omega)$, we define the cost functional $J : Q \times \mathcal{V} \to \mathbb{R}$ as

$$J(q,u) := \frac{1}{2}\|u - u_d\|_2^2 + \frac{\alpha}{2}\|q\|_2^2.$$

We aim to solve the following optimal control problem:

$$\text{Minimize } J(q,u) \quad \text{subject to (7.16) and } (q,u) \in \mathcal{Q}_{\text{ad}} \times \mathcal{V}. \tag{P}$$

We tacitly let $J(q,u) = \infty$ whenever $u \notin L^2(\Omega)$. For the finite element analysis of (P), we will later utilize the following relation, which holds for all $(q_1,u_1),(q_2,u_2) \in Q \times \mathcal{V}$ due to the parallelogram law:

$$\begin{aligned} &\frac{1}{2}\left\|\frac{u_1-u_2}{2}\right\|_2^2 + \frac{\alpha}{2}\left\|\frac{q_1-q_2}{2}\right\|_2^2 + J\left(\frac{1}{2}(q_1+q_2),\frac{1}{2}(u_1+u_2)\right) \\ &\quad \le \frac{1}{2}J(q_1,u_1) + \frac{1}{2}J(q_2,u_2). \end{aligned} \tag{7.19}$$

Further, we often make use of the continuous embedding $W^{1,p}(\Omega) \subset L^2(\Omega)$ for $p \ge \frac{2d}{d+2}$. As a start, we deal with the existence of solutions to (P).

**Theorem 7.3.3.** *For $p \in (1,\infty)$ and $\varepsilon \ge 0$, the optimal control problem (P) has at least one globally optimal control $\overline{q} \in \mathcal{Q}_{\text{ad}}$ with corresponding optimal state $\overline{u} = u(\overline{q}) \in \mathcal{V}$.*

*Proof.* The proof follows standard arguments; see [17, 31]. According to Lemma 7.3.1, for each control $q \in \mathcal{Q}_{\text{ad}}$, the state equation (7.16) has a unique solution $u = u(q) \in \mathcal{V}$. The functional $J$ is bounded from below. Thus there exists

$$j := \inf_{q \in \mathcal{Q}_{\text{ad}}} J(q,u(q)).$$

Let $\{(q_n,u_n)\}_{n=1}^{\infty}$ be a minimizing sequence, i. e.,

$$q_n \in \mathcal{Q}_{\text{ad}}, \quad u_n := u(q_n), \quad J(q_n,u_n) \to j \quad \text{as } n \to \infty.$$

As $\mathcal{Q}_{\text{ad}}$ is nonempty, convex, closed, and bounded in $L^{\max\{p',2\}}(\Omega)$, it is weakly sequentially compact. Hence there exists a subsequence, denoted again by $\{q_n\}_{n=1}^{\infty}$, that weakly converges in $L^{\max\{p',2\}}(\Omega)$ to a function $\overline{q} \in \mathcal{Q}_{\text{ad}}$,

$$q_n \rightharpoonup \overline{q} \quad \text{weakly in } L^{\max\{p',2\}}(\Omega)$$

and thus strongly in $W^{-1,p'}(\Omega)$. Then Poincaré's inequality and Lemma 7.3.2 imply

$$\|u_n - u(\overline{q})\|_{1,p} \lesssim \|\nabla u_n - \nabla u(\overline{q})\|_p \to 0 \quad (n \to \infty), \tag{7.20}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

where in the case $p < 2$, (7.17) was also used. If $p \geq \frac{2d}{d+2}$, then from (7.20) we directly obtain that $u_n \to u(\overline{q})$ strongly in $L^2(\Omega)$. If $p < \frac{2d}{d+2}$, then we proceed differently. Since

$$\|u_n\|_2^2 \lesssim \|u_n - u_d\|_2^2 + \|u_d\|_2^2 \leq 2J(q_n, u_n) + \|u_d\|_2^2 \leq C,$$

we have in addition for some function $\overline{u} \in L^2(\Omega)$ that

$$u_n \rightharpoonup \overline{u} \quad \text{weakly in } L^2(\Omega),$$

and due to $p < 2$, $u_n \rightharpoonup \overline{u}$ weakly in $L^p(\Omega)$. In view of (7.20), we have $u_n \to u(\overline{q})$ strongly in $L^p(\Omega)$. As the weak limit is unique, this yields $\overline{u} = u(\overline{q})$, and hence

$$(q_n, u_n) \rightharpoonup (\overline{q}, \overline{u}) = (\overline{q}, u(\overline{q})) \quad \text{weakly in } L^2(\Omega) \times L^2(\Omega).$$

The functional $J$ is convex and continuous on $L^2(\Omega) \times L^2(\Omega)$, so it is weakly lower-semicontinuous. Thus, from this weak convergence we conclude

$$j = \lim_{n\to\infty} J(q_n, u_n) \geq \liminf_{n\to\infty} J(q_n, u_n) \geq J(\overline{q}, \overline{u}) \geq j,$$

i. e., $\overline{q}$ is an optimal control. The proof is completed. □

## 7.4 Finite element discretization

In this section, we introduce the discretization of the optimal control problem (7.1). We assume that $\Omega$ is either a convex polygonal/polyhedral domain or a bounded convex domain with $C^2$-boundary, where for the latter case, we assume that $d = 2$. Let $\mathbb{T}_h$ be a shape regular decomposition of $\Omega$ into $d$-dimensional simplices such that if $\Omega$ is polyhedral, then $\overline{\Omega} = \bigcup_{K\in\mathbb{T}_h} \overline{K}$, or if $\Omega$ has a curved boundary, then the corner points of

$$\overline{\Omega_h} := \bigcup_{K\in\mathbb{T}_h} \overline{K} \subset \overline{\Omega}$$

belong to $\partial\Omega$. By $h_K$ we denote the diameter of a cell $K \in \mathbb{T}_h$, and by $\rho_K$ the supremum of diameters of inscribed balls. The mesh parameter $h$ represents the maximum diameter of the cells, i. e., $h := \max\{h_K;\, K \in \mathbb{T}_h\}$. We assume that $\mathbb{T}_h$ is nondegenerate (see [7]), i. e.,

$$\max_{K\in\mathbb{T}_h} \frac{h_K}{\rho_K} \leq \kappa_0 \quad \forall h. \tag{7.21}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

For $K \in \mathbb{T}_h$, we define the set of neighbors $N_K$ and the neighborhood $S_K$ by

$$\begin{aligned} N_K &:= \{K' \in \mathbb{T}_h : \overline{K'} \cap \overline{K} \neq \emptyset\}, \\ S_K &:= \operatorname{int} \bigcup_{K' \in N_K} \overline{K'}. \end{aligned} \tag{7.22}$$

The sets $S_K$ are open, bounded, and connected. The nondegeneracy (7.21) of $\mathbb{T}_h$ implies the following properties of $\mathbb{T}_h$, in which all constants are independent of $h$:

$$|S_K| \sim |K| \quad \text{for all } K \in \mathbb{T}_h \quad \text{and} \quad \#N_K \leq m_0 \quad \text{for some } m_0 \in \mathbb{N}. \tag{7.23}$$

If $\partial\Omega$ is curved, then we need to be able to estimate the integral over the part of the domain not covered by $\Omega_h$. As the treatment of curved boundaries is not the purpose of the paper, in this case, we restrict ourselves to space dimension $d = 2$ for ease of presentation.

To this end, we need the following:

**Lemma 7.4.1.** *Let $\Omega$ be a convex domain with $C^2$-boundary, and let $d = 2$. Then for $u \in W^{1,2}(\Omega)$, the integral of $|u|^2$ on the stripe $\Sigma_h := \overline{\Omega} \setminus \overline{\Omega_h}$ is bounded by*

$$\int_{\Sigma_h} |u(x)|^2 \, \mathrm{d}x \lesssim h^2 \|u\|^2_{1,2;\Omega}. \tag{7.24}$$

Its proof can be found, e. g., in the proof of [37, Satz 3.3].

Let $\mathcal{P}_m(K)$ be the set of polynomials on $K$ of degree less than or equal to $m$. For the discretization of the state equation, we employ the space $\mathcal{V}_h \subset W^{1,\infty}(\Omega)$ of linear finite elements on the triangulation $\mathbb{T}_h$,

$$\hat{\mathcal{V}}_h = \{u_h \in C(\overline{\Omega_h}) \mid u_h|_K \in \mathcal{P}_1(K) \text{ for all } K \in \mathbb{T}_h \text{ and } u_h|_{\partial\Omega_h} = 0\}.$$

If $\Omega$ is polyhedral, then $\Omega_h = \Omega$, and we set $\mathcal{V}_h = \hat{\mathcal{V}}_h$. Otherwise, we define $\mathcal{V}_h$ as the space of functions $u_h \in \hat{\mathcal{V}}_h$ extended to $\overline{\Omega}$ by setting $u_h = 0$ on the stripe $\Sigma_h := \overline{\Omega} \setminus \overline{\Omega_h}$. For the discretization of the controls, we study two approaches:

(a) Variational discretization: The control variable is not explicitly discretized.

(b) Piecewise constant controls on the family of triangulations $\{\mathbb{T}_h\}$ introduced for the discretization of the state variables:

$$\hat{Q}_h^0 = \{q_h : \overline{\Omega_h} \to \mathbb{R} \mid q_h|_K \in \mathcal{P}_0(K) \text{ for all } K \in \mathbb{T}_h\}.$$

If $\Omega$ is polyhedral, then we define $Q_h^0 = \hat{Q}_h^0$; otherwise,

$$Q_h^0 = \{q_h \in Q \mid q_h|_{\Omega_h} \in \hat{Q}_h^0\}.$$

The discrete admissible set is $\mathcal{Q}_{h,\mathrm{ad}}^0 := Q_h^0 \cap \mathcal{Q}_{\mathrm{ad}}$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Note that for the case of curved boundary, the functions in $Q_h^0$ are not restricted to a finite-dimensional set on the stripe $\Sigma_h = \overline{\Omega} \setminus \overline{\Omega_h}$. Hence this space is formally infinite-dimensional. However, due to the control cost $\frac{\alpha}{2}\|q\|_2^2$ and the fact that $q_h|_{\Sigma_h}$ has no influence on the discrete solution of (7.28), the optimal discrete control will satisfy $q_h|_{\Sigma_h} = 0$ due to our choice $0 \in [q_a, q_b]$. This means that an implementation can work on the finite-dimensional set $\hat{Q}_h^0$. In the subsequent analysis, $\mathcal{Q}_{h,\mathrm{ad}}$ equals either $\mathcal{Q}_{\mathrm{ad}}$ in case of variational discretization or $\mathcal{Q}_{h,\mathrm{ad}}^0$ in case of cellwise constant discretization.

Moreover, we introduce $\hat{\Pi}_h : L^1(\Omega_h) \to \hat{Q}_h^0$ as the natural extension of the $L^2$-projection, i. e., for $q \in L^1(\Omega_h)$, we define $\hat{\Pi}_h q \in \hat{Q}_h^0$ by

$$(\hat{\Pi}_h q, \varphi_h)_{\Omega_h} = (q, \varphi_h)_{\Omega_h} \quad \forall \varphi_h \in \hat{Q}_h^0. \tag{7.25}$$

If $\Omega$ is polyhedral, then we set $\Pi_h = \hat{\Pi}_h$. Otherwise, we define $\Pi_h : Q \to Q_h^0$ by $(\Pi_h q)|_{\Omega_h} = \hat{\Pi}_h q$ and $(\Pi_h q)|_{\Sigma_h} = q|_{\Sigma_h}$ for all $q \in Q$. It is well known that the operator $\Pi_h$ satisfies for any $\nu \in [1, \infty]$ the stability estimate

$$\|\Pi_h q\|_{\nu;\Omega_h} \le \|q\|_{\nu;\Omega_h} \quad \forall q \in L^\nu(\Omega_h). \tag{7.26}$$

From the stability property (7.26) we can derive an interpolation estimate for $\Pi_h$.

**Lemma 7.4.2.** *There exists a constant $c > 0$ independent of $h$ such that for all $q \in W^{m,\nu}(\Omega)$ with $m \in \{0, 1\}$ and $\nu \in (1, \infty)$, we have*

$$\|q - \Pi_h q\|_{-1,\nu} + h\|q - \Pi_h q\|_\nu \le c h^{m+1} \|q\|_{m,\nu}. \tag{7.27}$$

We omit the proof, as it is a standard consequence of orthogonality, the definition of the norms, and standard error estimates for quasi-interpolation operators noting that by the definition of $\Pi_h$ the boundary stripe $\Sigma_h$ induces no error.

The Galerkin approximation of (7.16) consists in replacing the Banach space $\mathcal{V}$ by the finite element space $\mathcal{V}_h$:

For a given control $q \in \mathcal{Q}_{\mathrm{ad}}$, find the discrete state $u_h = u_h(q) \in \mathcal{V}_h$ with

$$\big(\boldsymbol{S}(\nabla u_h), \nabla \varphi_h\big)_\Omega = (q, \varphi_h)_\Omega \quad \forall \varphi_h \in \mathcal{V}_h. \tag{7.28}$$

The existence of a unique solution $u_h$ to (7.28) and an a priori estimate for $u_h$ in $W^{1,p}(\Omega)$ follow by using similar arguments as in the continuous case.

**Lemma 7.4.3.** *For each $p \in (1, \infty)$, there exists a unique solution $u_h \in \mathcal{V}_h$ to* (7.28)*. This discrete solution satisfies the a priori estimate*

$$\|\nabla u_h\|_p \le c_1 \big(\|q\|_{-1,p'}^{\frac{1}{p-1}} + c_2 \varepsilon\big), \tag{7.29}$$

*where $c_1 > 0$ only depends on $\Omega$ and $p$, and $c_2 = 1$ if $p < 2$ and $c_2 = 0$ otherwise.*

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

The following lemma is a discrete version of Lemma 7.3.2.

**Lemma 7.4.4.** *For $p \in (1,\infty)$, let $u_{h1} = u_h(q_1) \in \mathcal{V}_h$ and $u_{h2} = u_h(q_2) \in \mathcal{V}_h$ be the solutions to the discrete equation* (7.28) *for the right-hand sides $q_1 \in \mathcal{Q}_{\mathrm{ad}}$ and $q_2 \in \mathcal{Q}_{\mathrm{ad}}$. Then there exist constants, only depending on $p$ and $\Omega$, such that*

$$\left\|\boldsymbol{F}(\nabla u_{h1}) - \boldsymbol{F}(\nabla u_{h2})\right\|_2 \lesssim \begin{cases} \|\varepsilon + |\nabla u_{h1}| + |\nabla u_{h2}|\|_p^{\frac{2-p}{2}} \|q_1 - q_2\|_{-1,p'} & \text{for } p \le 2, \\ \|q_1 - q_2\|_{-1,p'}^{\frac{p'}{2}} & \text{for } p \ge 2, \end{cases}$$

$$\|\nabla u_{h1} - \nabla u_{h2}\|_p \lesssim \begin{cases} \|\varepsilon + |\nabla u_{h1}| + |\nabla u_{h2}|\|_p^{2-p} \|q_1 - q_2\|_{-1,p'} & \text{for } p \le 2, \\ \|q_1 - q_2\|_{-1,p'}^{\frac{1}{p-1}} & \text{for } p \ge 2. \end{cases}$$

*Proof.* The proof follows along the same lines as the proof of Lemma 7.3.2 if the space $\mathcal{V}$ is replaced by $\mathcal{V}_h$. □

Now let us consider the discrete optimal control problem. The discrete analog to $(P)$ reads:

$$\text{Minimize } J(q_h, u_h) \quad \text{subject to (7.28) and } (q_h, u_h) \in \mathcal{Q}_{h,\mathrm{ad}} \times \mathcal{V}_h. \tag{$P_h$}$$

Following the same arguments used for the proof of Theorem 7.3.3, we can conclude the existence of a solution to $(P_h)$.

**Lemma 7.4.5.** *For each $h > 0$, there exists an optimal control $\overline{q}_h$ with corresponding optimal state $\overline{u}_h$ of the minimization problem $(P_h)$.*

## 7.5 Discrete optimality system

In this section, we are concerned with an optimality system for $(P_h)$, which can be utilized for practical computation of the discrete optimal solution. We will close this section with a discussion on the continuous optimality system. For ease of exposition, we restrict ourselves to the particular nonlinear operator $(7.2)_2$, i. e., for

$$a(u)(\varphi) := \left(\boldsymbol{S}(\nabla u), \nabla\varphi\right)_\Omega, \quad \boldsymbol{S}(\nabla u) = \left(\varepsilon^2 + |\nabla u|^2\right)^{\frac{p-2}{2}} \nabla u,$$

we consider the discrete variational formulation of the state equation (7.28),

$$a(u_h)(\varphi_h) = (q, \varphi_h)_\Omega \quad \forall \varphi_h \in \mathcal{V}_h. \tag{7.30}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

On the discrete level, it can be shown that the semilinear form $a$ is Gâteaux differentiable for each $\varepsilon > 0$ with Gâteaux derivative

$$\begin{aligned} a'(v_h)(w_h, \varphi_h) = &\int_\Omega \left(\varepsilon^2 + |\nabla v_h|^2\right)^{\frac{p-2}{2}} \nabla w_h \cdot \nabla \varphi_h \, \mathrm{d}x \\ &+ (p-2) \int_\Omega \left(\varepsilon^2 + |\nabla v_h|^2\right)^{\frac{p-4}{2}} (\nabla v_h \cdot \nabla w_h)(\nabla v_h \cdot \nabla \varphi_h) \, \mathrm{d}x. \end{aligned} \tag{7.31}$$

Since this will be crucial for this section, here we limit ourselves to $\varepsilon > 0$.

Now we can define the adjoint problem associated with (7.28): Find $z_h \in \mathcal{V}_h$ such that

$$a'(u_h)(\varphi_h, z_h) = (u_h - u_d, \varphi_h)_\Omega \quad \forall \varphi_h \in \mathcal{V}_h. \tag{7.32}$$

The next lemma concerns the unique solvability of the discrete adjoint problem.

**Lemma 7.5.1.** *Let $p \in (1, \infty)$, $\varepsilon > 0$, and $b \in W^{-1,\max\{p',2\}}(\Omega)$ with $p' = \frac{p}{p-1}$. For each $h > 0$, there exists a unique solution $z_h \in \mathcal{V}_h$ to*

$$a'(u_h)(\varphi_h, z_h) = \langle b, \varphi_h \rangle \quad \forall \varphi_h \in \mathcal{V}_h, \tag{7.33}$$

*where $u_h$ solves* (7.28)*. The adjoint solution $z_h$ satisfies the a priori estimate*

$$\|\nabla z_h\|_{\min\{p,2\}} \leq c \|b\|_{-1,\max\{p',2\}}, \tag{7.34}$$

*where the constant $c$ only depends on $p$, $\varepsilon$, $\Omega$, $q_a$, $q_b$.*

*Proof.* First, we prove that if there exists a solution $z_h$ to problem (7.33), then $z_h$ is uniquely determined. To this end, we assume that $z_h^1$ and $z_h^2$ are two functions satisfying (7.33). Setting $\xi_h := z_h^1 - z_h^2$, we observe that

$$a'(u_h)(\varphi_h, \xi_h) = 0 \quad \forall \varphi_h \in \mathcal{V}_h. \tag{7.35}$$

We recall that $u_h$ is uniformly bounded in $W^{1,p}(\Omega)$ by the data; see (7.29). In the case $p \leq 2$, we may estimate the quantity $a'(u_h)(\xi_h, \xi_h)$ as follows:

$$\begin{aligned} a'(u_h)(\xi_h, \xi_h) \overset{(7.31)}{\geq} &\int_\Omega \left(\varepsilon^2 + |\nabla u_h|^2\right)^{\frac{p-2}{2}} |\nabla \xi_h|^2 \, \mathrm{d}x \\ &+ (p-2) \int_\Omega \left(\varepsilon^2 + |\nabla u_h|^2\right)^{\frac{p-4}{2}} |\nabla u_h|^2 |\nabla \xi_h|^2 \, \mathrm{d}x \end{aligned}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$$\begin{aligned} &\geq (p-1)\int_\Omega \left(\varepsilon^2 + |\nabla u_h|^2\right)^{\frac{p-2}{2}} |\nabla \xi_h|^2 \,\mathrm{d}x \\ &\geq (p-1)\int_\Omega \left(\varepsilon + |\nabla u_h|\right)^{p-2} |\nabla \xi_h|^2 \,\mathrm{d}x. \end{aligned}$$

Using Hölder's inequality, $q(x) \in [q_a, q_b]$ a. e., and (7.29) for $p \leq 2$ and $\varepsilon > 0$, we arrive at

$$a'(u_h)(\xi_h, \xi_h) \geq (p-1)\left\|\varepsilon + |\nabla u_h|\right\|_p^{p-2} \|\nabla \xi_h\|_p^2 \geq c\|\nabla \xi_h\|_p^2.$$

In the case $p > 2$, we can bound the quantity $a'(u_h)(\xi_h, \xi_h)$ from below as follows:

$$\begin{aligned} a'(u_h)(\xi_h, \xi_h) &\overset{(7.31)}{=} \int_\Omega \left(\varepsilon^2 + |\nabla u_h|^2\right)^{\frac{p-2}{2}} \nabla \xi_h \cdot \nabla \xi_h \,\mathrm{d}x \\ &\quad + (p-2)\int_\Omega \left(\varepsilon^2 + |\nabla u_h|^2\right)^{\frac{p-4}{2}} (\nabla u_h \cdot \nabla \xi_h)(\nabla u_h \cdot \nabla \xi_h) \,\mathrm{d}x \\ &\geq \int_\Omega \left(\varepsilon^2 + |\nabla u_h|^2\right)^{\frac{p-2}{2}} |\nabla \xi_h|^2 \,\mathrm{d}x \geq \int_\Omega \varepsilon^{p-2} |\nabla \xi_h|^2 \,\mathrm{d}x = \varepsilon^{p-2}\|\nabla \xi_h\|_2^2. \end{aligned}$$

To sum up, we may deduce that there exists a constant $c = c(p, \varepsilon, \Omega, q_a, q_b)$ with

$$a'(u_h)(\xi_h, \xi_h) \geq c\|\nabla \xi_h\|_{\min\{p,2\}}^2. \tag{7.36}$$

From (7.35), (7.36), and Poincaré's inequality we infer $\xi_h \equiv 0$, and hence $z_h^1 = z_h^2$. Since system (7.33) is linear and the space $\mathcal{V}_h$ is finite-dimensional, we can conclude from the uniqueness that there exists a solution $z_h$. For the proof of (7.34), we test (7.33) with $\varphi_h := z_h$. Then we can apply the same arguments that led to (7.36) to obtain

$$\|b\|_{-1,\max\{p',2\}} \|z_h\|_{1,\min\{p,2\}} \geq \langle b, z_h \rangle = a'(u_h)(z_h, z_h) \geq \|\nabla z_h\|_{\min\{p,2\}}^2.$$

Together with Poincaré's inequality, this yields the statement. □

With the help of the discrete adjoint state, we can now formulate an optimality system for $(P_h)$:

**Lemma 7.5.2.** *Let $\varepsilon > 0$. If a control $\overline{q}_h \in \mathcal{Q}_{h,\mathrm{ad}}$ with state $\overline{u}_h = u_h(\overline{q}_h) \in \mathcal{V}_h$ is an optimal solution to problem $(P_h)$, then there exists an adjoint state $\overline{z}_h \in \mathcal{V}_h$, so that*

$$a(\overline{u}_h)(\varphi_h) = (\overline{q}_h, \varphi_h)_\Omega \quad \forall \varphi_h \in \mathcal{V}_h, \tag{7.37a}$$

$$a'(\overline{u}_h)(\overline{z}_h, \varphi_h) = (\overline{u}_h - u_d, \varphi_h)_\Omega \quad \forall \varphi_h \in \mathcal{V}_h, \tag{7.37b}$$

$$(\alpha \overline{q}_h + \overline{z}_h, \delta q_h - \overline{q}_h)_\Omega \geq 0 \quad \forall \delta q_h \in \mathcal{Q}_{h,\mathrm{ad}}. \tag{7.37c}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Remark 7.5.3.** It is well known that the variational inequality (7.37c) has a pointwise almost everywhere representation; see, e. g., [40]. Indeed, (7.37c) can be rewritten using the projection $P_{[q_a,q_b]}$ onto the interval $[q_a, q_b]$ defined by

$$P_{[q_a,q_b]}(f(x)) = \min(q_b, \max(q_a, f(x))). \tag{7.38}$$

In the case $\mathcal{Q}_{h,\mathrm{ad}} = \mathcal{Q}_{\mathrm{ad}}$, a control $\overline{q}_h \in \mathcal{Q}_{h,\mathrm{ad}}$ solving $(P_h)$ necessarily satisfies (7.37), and thus the control $\overline{q}_h$ and the solution $\overline{z}_h$ of (7.37b) satisfy the projection formula

$$\overline{q}_h = P_{[q_a,q_b]}\left(-\frac{1}{\alpha}\overline{z}_h\right).$$

In the case $\mathcal{Q}_{h,\mathrm{ad}} = \mathcal{Q}^0_{h,\mathrm{ad}}$, we have

$$\overline{q}_h = P_{[q_a,q_b]}\left(-\frac{1}{\alpha}\Pi_h\overline{z}_h\right),$$

where $\Pi_h$ denotes the $L^2$-projection on $\mathcal{Q}^0_{h,\mathrm{ad}}$.

The question arises whether an analogous optimality system for $(P)$ can be formulated. A closer look at (7.31) however reveals that this is not an easy task: The natural regularity for the state $u \in W^{1,p}(\Omega)$ is not sufficient to formulate a well-defined, invertible, Gâteaux derivative, as already discussed in the introduction.

Still, we can pass to the limit in the discrete adjoint. According to the a priori estimate (7.34) and Lemma 7.4.3, the discrete adjoint solution $\overline{z}_h$ is uniformly bounded in $W^{1,\min\{p,2\}}(\Omega)$ for $p \geq \frac{2d}{d+2}$ as it is shown by the following calculation:

$$\begin{aligned}\|\nabla\overline{z}_h\|_{\min\{p,2\}} &\leq c\|u_h - u_d\|_{-1,\max\{p',2\}} = \sup_{\varphi\in W_0^{1,\min\{p,2\}}(\Omega)} \frac{(u_h - u_d, \varphi)_\Omega}{\|\varphi\|_{1,\min\{p,2\}}} \\ &\leq \sup_{\varphi\in W_0^{1,\min\{p,2\}}(\Omega)} \frac{\|u_h - u_d\|_2\|\varphi\|_2}{\|\varphi\|_{1,\min\{p,2\}}} \lesssim \|u_h\|_{1,p} + \|u_d\|_2 \lesssim C.\end{aligned}$$

Hence there exists a function $\overline{z} \in W_0^{1,\min\{p,2\}}(\Omega)$ such that, up to a subsequence,

$$\overline{z}_h \rightharpoonup \overline{z} \quad \text{weakly in } W_0^{1,\min\{p,2\}}(\Omega) \quad (h \to 0). \tag{7.39}$$

Due to the compact embedding $W^{1,\min\{p,2\}}(\Omega) \subset L^2(\Omega)$ for $p > \frac{2d}{d+2}$, we get

$$\overline{z}_h \to \overline{z} \quad \text{strongly in } L^2(\Omega) \quad (h \to 0). \tag{7.40}$$

Consequently, the projection formula yields strong convergence of the controls

$$\overline{q}_h = P_{[q_a,q_b]}\left(-\frac{1}{\alpha}\overline{z}_h\right) \to \overline{q} = P_{[q_a,q_b]}\left(-\frac{1}{\alpha}\overline{z}\right) \quad \text{in } L^2(\Omega) \quad (h \to 0)$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

for $p > \frac{2d}{d+2}$ in the case of variational discretization; see Remark 7.5.3. This also shows the additional regularity $q \in W^{1,\min\{p,2\}}(\Omega)$ for any such limit point.

## 7.6 FE approximation of the *p*-Laplace equation

Before analyzing the convergence of the discretized optimal control problem, we collect and extend several results regarding the FE approximation of the $p$-Laplace equation. The first lemma states that the Galerkin approximation is a quasi-best-approximation with respect to the natural distance.

**Lemma 7.6.1** (Best-approximation in quasinorms). *For $\varepsilon \geq 0$ and $p \in (1,\infty)$, let $u \in \mathcal{V}$ be the unique solution of* (7.16)*, and let $u_h \in \mathcal{V}_h$ its finite element approximation, i. e., $u_h \in \mathcal{V}_h$ is the unique solution of* (7.28)*. Then, for $\Sigma_h := \overline{\Omega} \setminus \overline{\Omega_h}$,*

$$\left\|\boldsymbol{F}(\nabla u) - \boldsymbol{F}(\nabla u_h)\right\|_{2;\Omega} \lesssim \inf_{\varphi_h \in \mathcal{V}_h} \left\|\boldsymbol{F}(\nabla u) - \boldsymbol{F}(\nabla \varphi_h)\right\|_{2;\Omega}, \tag{7.41a}$$

$$\left\|\boldsymbol{F}(\nabla u) - \boldsymbol{F}(\nabla u_h)\right\|_{2;\Omega} \lesssim \inf_{\varphi_h \in \mathcal{V}_h} \left\|\boldsymbol{F}(\nabla u) - \boldsymbol{F}(\nabla \varphi_h)\right\|_{2;\Omega_h} + \left\|\boldsymbol{F}(\nabla u)\right\|_{2;\Sigma_h}, \tag{7.41b}$$

*where the constants only depend on p (they are independent of h and ε).*

*Proof.* For polyhedral $\Omega$, the lemma is proven in [22]. Using Lemma 7.2.2 and the Galerkin orthogonality ($\mathcal{V}_h \subset \mathcal{V}$), we can deduce that for arbitrary $\varphi_h \in \mathcal{V}_h$,

$$\begin{aligned}
\int_\Omega \varphi_{|\nabla u|}(|\nabla u - \nabla u_h|)\,\mathrm{d}x &\sim \int_\Omega \left(\boldsymbol{S}(\nabla u) - \boldsymbol{S}(\nabla u_h)\right)\cdot(\nabla u - \nabla u_h)\,\mathrm{d}x \\
&\sim \int_\Omega \left(\boldsymbol{S}(\nabla u) - \boldsymbol{S}(\nabla u_h)\right)\cdot(\nabla u - \nabla \varphi_h)\,\mathrm{d}x \\
&\lesssim \int_\Omega \varphi'_{|\nabla u|}(|\nabla u - \nabla u_h|)|\nabla u - \nabla \varphi_h|\,\mathrm{d}x.
\end{aligned}$$

Applying Young's inequality (7.12) to the shifted function $\varphi_{|\nabla u|}$, we obtain

$$\int_\Omega \varphi_{|\nabla u|}(|\nabla u - \nabla u_h|)\,\mathrm{d}x \lesssim \int_\Omega \varphi_{|\nabla u|}(|\nabla u - \nabla \varphi_h|)\,\mathrm{d}x.$$

Using Lemma 7.2.2 and taking the infimum over all $\varphi_h \in \mathcal{V}_h$, we arrive at statement (7.41a). We have $\varphi_h|_{\Sigma_h} = 0$ for all $\varphi_h \in \mathcal{V}_h$, and thus (7.41b) easily follows from inequality (7.41a). □

The Scott–Zhang interpolation operator $j_h : W_0^{1,1}(\Omega) \to \mathcal{V}_h$ (see [39]) is defined so that it fulfills $j_h v = v$ for all $v \in \mathcal{V}_h$ and preserves homogeneous boundary conditions.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

It is also suitable for interpolation in quasi-norms as it satisfies the following property (see [22]): For all $v \in W^{1,p}(\Omega)$ and $K \in \mathbb{T}_h$, we have

$$\fint_K \left|\boldsymbol{F}(\nabla v) - \boldsymbol{F}(\nabla j_h v)\right|^2 \,\mathrm{d}x \lesssim \inf_{\boldsymbol{\eta}\in\mathbb{R}^d} \fint_{S_K} \left|\boldsymbol{F}(\nabla v) - \boldsymbol{F}(\boldsymbol{\eta})\right|^2 \,\mathrm{d}x, \tag{7.42}$$

where the constant only depends on $p$. In particular, it is independent of $h$ and $\varepsilon$. On the basis of (7.42), it is a simple matter to derive an interpolation estimate in quasi-norms: As the function $\boldsymbol{F}$ is surjective, (7.42) implies

$$\fint_K \left|\boldsymbol{F}(\nabla v) - \boldsymbol{F}(\nabla j_h v)\right|^2 \,\mathrm{d}x \lesssim \inf_{\boldsymbol{\xi}\in\mathbb{R}^d} \fint_{S_K} \left|\boldsymbol{F}(\nabla v) - \boldsymbol{\xi}\right|^2 \,\mathrm{d}x. \tag{7.43}$$

Now let us assume that $v \in W^{1,p}(\Omega)$ satisfies the regularity $\boldsymbol{F}(\nabla v) \in W^{1,2}(\Omega)^d$. If we choose $\boldsymbol{\xi} = \langle \boldsymbol{F}(\nabla v)\rangle_{S_K}$ in (7.43), then we can apply Poincaré's inequality (7.4):

$$\fint_K \left|\boldsymbol{F}(\nabla v) - \boldsymbol{F}(\nabla j_h v)\right|^2 \,\mathrm{d}x \lesssim \fint_{S_K} h_K^2 \left|\nabla \boldsymbol{F}(\nabla v)\right|^2 \,\mathrm{d}x. \tag{7.44}$$

To obtain a global version of (7.44), we sum inequality (7.44) over all elements $K \in \mathbb{T}_h$ and use the mesh properties (7.23):

$$\left\|\boldsymbol{F}(\nabla v) - \boldsymbol{F}(\nabla j_h v)\right\|_{2;\Omega_h} \le ch \left\|\nabla \boldsymbol{F}(\nabla v)\right\|_{2;\Omega_h}. \tag{7.45}$$

Combining Lemma 7.6.1 and (7.45), we obtain an error estimate in quasi-norms.

**Lemma 7.6.2.** *For $\varepsilon \ge 0$ and $p \in (1,\infty)$, let $u \in \mathcal{V}$ be the unique solution of* (7.16)*, and let $u_h \in \mathcal{V}_h$ be its finite element approximation, i. e., $u_h \in \mathcal{V}_h$ is the unique solution of* (7.28)*. In case of curved $\partial\Omega$, we require that $d = 2$.*

(i) *If $u$ satisfies the regularity assumption $\boldsymbol{F}(\nabla u) \in W^{1,2}(\Omega)^d$, then*

$$\left\|\boldsymbol{F}(\nabla u) - \boldsymbol{F}(\nabla u_h)\right\|_2 \le ch \left\|\boldsymbol{F}(\nabla u)\right\|_{1,2}.$$

(ii) *If $u$ satisfies $\boldsymbol{S}(\nabla u) \in W^{1,2}(\Omega)^d$ and $u \in W^{2,2}(\Omega)$, then*

$$\left\|\boldsymbol{F}(\nabla u) - \boldsymbol{F}(\nabla u_h)\right\|_2 \le ch \left\|\boldsymbol{S}(\nabla u)\right\|_{1,2}^{\frac{1}{2}} \|u\|_{2,2}^{\frac{1}{2}}.$$

*All constants $c$ only depend on $p$ (they are independent of $h$ and $\varepsilon$).*

*Proof.* (i) For polyhedral $\Omega$, the statement directly follows by combining (7.41a) and (7.45). If $\partial\Omega$ is curved, then we use (7.41b) and (7.45) and take into account

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Lemma 7.4.1:

$$\begin{aligned}\left\|\boldsymbol{F}(\nabla u)-\boldsymbol{F}(\nabla u_h)\right\|_{2;\Omega} &\overset{(7.41b)}{\lesssim} \inf_{\varphi_h\in\mathcal{V}_h}\left\|\boldsymbol{F}(\nabla u)-\boldsymbol{F}(\nabla\varphi_h)\right\|_{2;\Omega_h}+\left\|\boldsymbol{F}(\nabla u)\right\|_{2;\Sigma_h}\\ &\lesssim \left\|\boldsymbol{F}(\nabla u)-\boldsymbol{F}(\nabla j_h u)\right\|_{2;\Omega_h}+\left\|\boldsymbol{F}(\nabla u)\right\|_{2;\Sigma_h} \overset{(7.45),\,(7.24)}{\lesssim} ch\left\|\boldsymbol{F}(\nabla u)\right\|_{1,2;\Omega}.\end{aligned}$$

(ii) From (7.45), the pointwise estimate with constants independent of $\varepsilon$ (cf. [4])

$$\left|\nabla\boldsymbol{F}(\nabla v)\right|^2 \sim \left|\nabla\boldsymbol{S}(\nabla v)\right|\left|\nabla^2 v\right|,$$

and the Hölder inequality, we get that for all $v\in W^{2,2}(\Omega)$ with $\boldsymbol{S}(\nabla v)\in W^{1,2}(\Omega)^d$,

$$\left\|\boldsymbol{F}(\nabla v)-\boldsymbol{F}(\nabla j_h v)\right\|_{2;\Omega_h} \lesssim h\left\|\nabla\boldsymbol{S}(\nabla v)\right\|_{2;\Omega_h}^{\frac{1}{2}}\left\|\nabla^2 v\right\|_{2;\Omega_h}^{\frac{1}{2}}. \tag{7.46}$$

If $\Omega$ is polyhedral, then the statement directly follows by combining (7.41a) and (7.46). If $\partial\Omega$ is curved, then we use (7.41b), (7.46), Lemma 7.2.2, and Lemma 7.4.1:

$$\begin{aligned}\left\|\boldsymbol{F}(\nabla u)-\boldsymbol{F}(\nabla u_h)\right\|_{2;\Omega}^2 &\lesssim \inf_{\varphi_h\in\mathcal{V}_h}\left\|\boldsymbol{F}(\nabla u)-\boldsymbol{F}(\nabla\varphi_h)\right\|_{2;\Omega_h}^2+\left\|\boldsymbol{F}(\nabla u)\right\|_{2;\Sigma_h}^2\\ &\lesssim \left\|\boldsymbol{F}(\nabla u)-\boldsymbol{F}(\nabla j_h u)\right\|_{2;\Omega_h}^2+\left(\boldsymbol{S}(\nabla u),\nabla u\right)_{\Sigma_h}\\ &\lesssim h^2\left\|\nabla\boldsymbol{S}(\nabla u)\right\|_{2;\Omega_h}\left\|\nabla^2 u\right\|_{2;\Omega_h}+\left\|\boldsymbol{S}(\nabla u)\right\|_{2;\Sigma_h}\left\|\nabla u\right\|_{2;\Sigma_h}\\ &\lesssim h^2\left\|\nabla\boldsymbol{S}(\nabla u)\right\|_{2;\Omega_h}\left\|\nabla^2 u\right\|_{2;\Omega_h}+h\left\|\boldsymbol{S}(\nabla u)\right\|_{1,2;\Omega}\cdot h\|u\|_{2,2;\Omega}.\end{aligned}$$

Taking the square root, we obtain the stated a priori estimate. □

If the solution $u$ of the state equation satisfies the regularity assumption (7.15), according to Lemma 7.6.2(ii), the error measured in the natural distance can be bounded in terms of the control $q$. From this we get error bounds in the $W^{1,p}$-norm.

**Corollary 7.6.3.** *For $p\in(1,\infty)$, $\varepsilon\geq 0$, and any $q\in\mathcal{Q}_{\mathrm{ad}}$, let $u=u(q)\in\mathcal{V}$ be the solution of* (7.16)*, and let $u_h=u_h(q)\in\mathcal{V}_h$ be its discrete approximation, i. e., the solution of* (7.28)*. In case of curved $\partial\Omega$, we require that $d=2$. If $u$ satisfies Assumption 7.2.4, there exist constants only depending on $p$ with*

$$\left\|\boldsymbol{F}(\nabla u)-\boldsymbol{F}(\nabla u_h)\right\|_2\leq ch,\quad \|\nabla u-\nabla u_h\|_p\leq\begin{cases}ch & \text{for } p\leq 2,\\ ch^{\frac{2}{p}} & \text{for } p\geq 2.\end{cases} \tag{7.47}$$

*In particular, the constants do not depend on the mesh size $h$ and $\varepsilon$.*

*Proof.* The error estimate in the natural distance directly follows from Lemma 7.6.2(ii), Assumption 7.2.4, and the inequality $\|q\|_{\max\{2,p'\}}\lesssim\max\{|q_a|,|q_b|\}$:

$$\left\|\boldsymbol{F}(\nabla u)-\boldsymbol{F}(\nabla u_h)\right\|_2\leq Ch\|q\|_2^{\frac{1}{2}}\|q\|_{\max\{2,p'\}}^{\frac{\gamma}{2}}\leq Ch.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

To derive the error estimates in the $W^{1,p}$-norm, we apply Lemma 7.2.3:

$$\left\|\boldsymbol{F}(\nabla u) - \boldsymbol{F}(\nabla u_h)\right\|_2^2 \gtrsim \begin{cases} \|\varepsilon + |\nabla u| + |\nabla u_h|\|_p^{p-2} \|\nabla u - \nabla u_h\|_p^2 & \text{for } p \le 2, \\ \|\nabla u - \nabla u_h\|_p^p & \text{for } p \ge 2, \end{cases}$$

and, if $p \le 2$, then use the stability Lemmas 7.3.1 and 7.4.3 and the inequality $\|q\|_{p'} \lesssim \max\{|q_a|, |q_b|\}$. □

If higher regularity is not available for the solution of state equation (7.16), then we still have the strong convergence of its finite element approximation in $W^{1,p}(\Omega)$.

**Lemma 7.6.4.** *For $p \in (1,\infty)$ and $\varepsilon \ge 0$, let $u \in \mathcal{V}$ be the unique solution of the state equation* (7.16)*, and let $u_h \in \mathcal{V}_h$ be the unique solution of its discrete approximation* (7.28)*, each for the right-hand side $q \in W^{-1,p'}(\Omega)$. Then $u_h$ converges strongly in $\mathcal{V}$ to $u$ as $h \to 0$, i. e.,*

$$\lim_{h\to 0} \|u - u_h\|_{1,p} = 0. \tag{7.48}$$

*Proof.* The lemma is proven in [12] for the case of polyhedral Ω.. Let Ω be a bounded convex domain with $\partial\Omega \in C^2$. Since $C_0^\infty(\Omega)$ is dense in $\mathcal{V}$, there exists a sequence $(\Phi_n) \subset C_0^\infty(\Omega)$ such that

$$\|u - \Phi_n\|_{1,p} \to 0 \quad (n \to \infty). \tag{7.49}$$

Let $i_h : C(\overline{\Omega}) \to \mathcal{V}_h$ denote the Lagrange interpolation operator; see [15]. On the stripe $\Sigma_h = \overline{\Omega} \setminus \overline{\Omega_h}$, we can set $(i_h\Phi)|_{\Sigma_h} = 0$ for $\Phi \in C(\overline{\Omega})$. Applied to $\Phi_n$, for all $K \in \mathbb{T}_h$, it satisfies

$$\|\Phi_n - i_h\Phi_n\|_{1,p;K} \le c|K|^{1/p} h_K \|\Phi_n\|_{2,\infty;K}. \tag{7.50}$$

Because of Lemma 7.6.1, the finite element solution $u_h$ fulfills

$$\begin{aligned} \left\|\boldsymbol{F}(\nabla u) - \boldsymbol{F}(\nabla u_h)\right\|_2 &\lesssim \inf_{\varphi_h \in \mathcal{V}_h} \left\|\boldsymbol{F}(\nabla u) - \boldsymbol{F}(\nabla \varphi_h)\right\|_2 \lesssim \left\|\boldsymbol{F}(\nabla u) - \boldsymbol{F}(\nabla i_h\Phi_n)\right\|_2 \\ &\lesssim \left\|\boldsymbol{F}(\nabla u) - \boldsymbol{F}(\nabla \Phi_n)\right\|_2 + \left\|\boldsymbol{F}(\nabla \Phi_n) - \boldsymbol{F}(\nabla i_h\Phi_n)\right\|_2. \end{aligned} \tag{7.51}$$

As the support of $\Phi_n$, $\operatorname{supp}(\Phi_n)$, is compact and $\operatorname{supp}(\Phi_n) \subset \Omega$, there exists $h_0 = h_0(n) > 0$ such that

$$h < h_0(n) \quad \Rightarrow \quad \operatorname{supp}(\Phi_n) \subset \overline{\Omega_h} \quad \text{with} \quad \overline{\Omega_h} = \bigcup_{K\in\mathbb{T}_h} \overline{K}.$$

We can then infer from (7.50) that for $h < h_0(n)$, we have the estimate

$$\|\Phi_n - i_h\Phi_n\|_{1,p;\Omega} \le c|\Omega|^{1/p} h \|\Phi_n\|_{2,\infty;\Omega}. \tag{7.52}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

As the natural distance relates to the $W_0^{1,p}$-norm (see Lemma 7.2.3), we can combine (7.51) and (7.52) to obtain, for each $n \in \mathbb{N}$,

$$\lim_{h\to 0} \|\boldsymbol{F}(\nabla u) - \boldsymbol{F}(\nabla u_h)\|_2 \lesssim \|\boldsymbol{F}(\nabla u) - \boldsymbol{F}(\nabla \Phi_n)\|_2.$$

Employing Lemma 7.2.3 and recalling (7.49), from this we infer the statement. □

## 7.7 Convergence of the approximation of the optimal control problem

This section contains the main results of the paper: Without assuming any regularity, for the case of piecewise constant controls, we show that the sequence of discrete global optimal solutions $(\overline{q}_h, \overline{u}_h)$ has a strong accumulation point $(\overline{q}, \overline{u}) \in \mathcal{Q}_{\mathrm{ad}} \times \mathcal{V}$ that is a global optimal solution of the original optimal control problem. Under the regularity Assumption 7.2.4, we then prove a priori error estimates quantifying the order of convergence for both variational discretization and piecewise constant controls.

**Theorem 7.7.1** (Convergence of global minimizers). *For $\varepsilon \in [0, \infty)$ and $p \in [\frac{2d}{d+2}, \infty)$, let $\boldsymbol{S}$ satisfy Assumption 7.2.1. For each $h > 0$, let $\overline{q}_h \in \mathcal{Q}_{h,\mathrm{ad}}$ be a discrete global optimal control, and let $\overline{u}_h = u_h(\overline{q}_h) \in \mathcal{V}_h$ be the corresponding discrete optimal state, i. e., $(\overline{q}_h, \overline{u}_h)$ solves $(P_h)$. Then the sequence $(\overline{q}_h, \overline{u}_h)$ has a weak accumulation point $(\overline{q}, \overline{u}) \in \mathcal{Q}_{\mathrm{ad}} \times \mathcal{V}$. Further, any weak accumulation point is also a strong accumulation point, i. e., up to a subsequence,*

$$\overline{q}_h \to \overline{q} \quad \text{in } L^2(\Omega), \quad \overline{u}_h \to \overline{u} \quad \text{in } W^{1,p}(\Omega) \quad (h \to 0).$$

*Moreover, any such point $(\overline{q}, \overline{u})$ is a global optimal solution of $(P)$.*

*Proof.* For each $h > 0$, let $(\overline{q}_h, \overline{u}_h)$ be a global solution of $(P_h)$. Weak accumulation points of $(\overline{q}_h, \overline{u}_h)$ exist in $L^{\max\{p',2\}}(\Omega) \times W_0^{1,p}(\Omega)$ due to the uniform a priori bounds

$$\|\overline{q}_h\|_{\max\{p',2\}} \lesssim \max\{|q_a|, |q_b|\}, \quad \|\overline{u}_h\|_{1,p} \overset{(7.29)}{\lesssim} C \quad \text{uniformly in } h. \tag{7.53}$$

Now let $(\bar{q}, \bar{u})$ be a global minimizer of $(P)$, whose existence is ensured by Theorem 7.3.3.

For piecewise constant controls, i. e., $\mathcal{Q}_{h,\mathrm{ad}} = \mathcal{Q}_{h,\mathrm{ad}}^0$, we define $\Pi_h$ as the $L^2$-projection given in Section 7.4. For variational controls, i. e., $\mathcal{Q}_{h,\mathrm{ad}} = \mathcal{Q}_{\mathrm{ad}}$, we set $\Pi_h = \mathrm{Id}$. Then the sequence of minimizers $(\overline{q}_h, \overline{u}_h)$ satisfies

$$J(\overline{q}_h, \overline{u}_h) \le J\big(\Pi_h \overline{q}, u_h(\Pi_h \overline{q})\big). \tag{7.54}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Using $W^{1,p}(\Omega) \subset L^2(\Omega)$ for $p \geq \frac{2d}{d+2}$, we can then infer

$$\|u_h(\Pi_h\overline{q}) - u(\overline{q})\|_2 \leq \underbrace{\|u_h(\Pi_h\overline{q}) - u_h(\overline{q})\|_2}_{\to 0 \text{ due to Lemma 7.4.4, (7.29), (7.27)}} + \underbrace{\|u_h(\overline{q}) - u(\overline{q})\|_2}_{\to 0 \text{ due to Lemma 7.6.4}} \overset{h\to 0}{\longrightarrow} 0.$$

Therefore inequality (7.54) implies

$$\limsup_{h\to 0} J(\overline{q}_h, \overline{u}_h) \leq \limsup_{h\to 0} J(\Pi_h\overline{q}, u_h(\Pi_h\overline{q})) = J(\overline{q}, u(\overline{q})).$$

Hence any weak limit $(q,u)$ of $(\overline{q}_h, \overline{u}_h)$ in $L^{\max\{p',2\}}(\Omega) \times W^{1,p}(\Omega)$ satisfies

$$J(q,u) \leq \liminf_{h\to 0} J(\overline{q}_h, \overline{u}_h) \leq \limsup_{h\to 0} J(\overline{q}_h, \overline{u}_h) \leq J(\bar{q}, \bar{u}),$$

since $J$ is weakly lower semicontinuous. Analogously to the proof of Theorem 7.3.3, we can show that $(q,u)$ solves (7.16). Hence $(q,u)$ is a global minimizer of $(P)$, and

$$J(\overline{q}_h, \overline{u}_h) \to J(q,u) \quad (h \to 0). \tag{7.55}$$

Further, $\overline{q}_h \rightharpoonup q$ weakly in $L^{\max\{p',2\}}(\Omega)$ implies that $\overline{q}_h \to q$ strongly in $W^{-1,p'}(\Omega)$. We apply Lemmas 7.6.4 and 7.4.4 (together with the bound (7.29) in the case $p \leq 2$) to see that

$$\|\overline{u}_h - u\|_{1,p} \leq \underbrace{\|u_h(\overline{q}_h) - u_h(q)\|_{1,p}}_{\to 0 \text{ due to Lemma 7.4.4, (7.29)}} + \underbrace{\|u_h(q) - u(q)\|_{1,p}}_{\to 0 \text{ due to Lemma 7.6.4}} \overset{h\to 0}{\longrightarrow} 0$$

and thus to obtain the strong convergence $\overline{u}_h \to u$ in $W_0^{1,p}(\Omega)$. By the parallelogram law (7.19), for $\hat{q} = \frac{1}{2}(\overline{q}_h + q)$ and $\hat{u} = \frac{1}{2}(\overline{u}_h + u)$, we get

$$\frac{1}{8}\|\overline{u}_h - u\|^2 + \frac{\alpha}{8}\|\overline{q}_h - q\|_2^2 \leq \frac{1}{2}J(q,u) + \frac{1}{2}J(\overline{q}_h, \overline{u}_h) - J(\hat{q}, \hat{u}).$$

We set $\tilde{u} = u(\hat{q})$. Then $J(q,u) \leq J(\hat{q}, \tilde{u})$, and hence

$$\frac{1}{8}\|\overline{u}_h - u\|^2 + \frac{\alpha}{8}\|\overline{q}_h - q\|_2^2 \leq \frac{1}{2}J(\overline{q}_h, \overline{u}_h) - \frac{1}{2}J(q,u) + \big(J(\hat{q}, \tilde{u}) - J(\hat{q}, \hat{u})\big). \tag{7.56}$$

In view of (7.55), the first difference on the right-hand side of (7.56) goes to zero as $h \to 0$. For the last sum, in parenthesis, we notice that

$$2J(\hat{q}, \tilde{u}) - 2J(\hat{q}, \hat{u}) = \|\tilde{u} - u_d\|_2^2 - \|\hat{u} - u_d\|_2^2. \tag{7.57}$$

As already shown, $\overline{u}_h \to u$ in $W^{1,p}(\Omega)$, and hence $\hat{u} \to u$ in $W^{1,p}(\Omega)$. Moreover, as $\overline{q}_h \to q$ in $W^{-1,p'}(\Omega)$, we have $\hat{q} \to q$ in $W^{-1,p'}(\Omega)$, and thus by Lemma 7.3.2

$$\tilde{u} = u(\hat{q}) \to u = u(q) \quad \text{strongly in } W^{1,p}(\Omega).$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

By our assumption on $p$, $p \geq \frac{2d}{d+2}$, we therefore obtain

$$\tilde{u} \to u \quad \text{strongly in } L^2(\Omega), \quad \hat{u} \to u \quad \text{strongly in } L^2(\Omega).$$

Combining this, (7.57), (7.55), and (7.56), we conclude that $\overline{q}_h \to q$ strongly in $L^2(\Omega)$. □

To prove the rates of convergence, we follow the approach presented in [36]. First, let us deal with the variational discretization, i. e., only the state space is discretized. For brevity of presentation, we name this problem

$$\text{Minimize } J(q_h, u_h) \quad \text{subject to (7.28) and } (q_h, u_h) \in \mathcal{Q}_{\text{ad}} \times \mathcal{V}_h. \tag{$P_s$}$$

**Theorem 7.7.2** (Convergence rates for variational discretization). *For $\varepsilon \in [0, \infty)$ and $p \in [\frac{2d}{d+2}, \infty)$, let Assumptions 7.2.1 and 7.2.4 be satisfied. For each $h > 0$, let $(\overline{q}_h, \overline{u}_h) \in \mathcal{Q}_{\text{ad}} \times \mathcal{V}_h$ be a global solution of the semidiscretized optimization problem $(P_s)$, and let $(\overline{q}, \overline{u}) \in \mathcal{Q}_{\text{ad}} \times \mathcal{V}$ be a global solution of $(P)$. Then there exists a constant $c > 0$ independent of $h$ such that*

$$\left|J(\overline{q}, \overline{u}) - J(\overline{q}_h, \overline{u}_h)\right| \leq c h^{\min\{1, \frac{2}{p}\}}. \tag{7.58}$$

*Proof.* We define $u_h \in \mathcal{V}_h$ as the solution to (7.28) for the control $\overline{q}$. For $p \in (1, \infty)$, Corollary 7.6.3 provides us with the estimate

$$\|\nabla \overline{u} - \nabla u_h\|_p \leq c h^{\min\{1, \frac{2}{p}\}}.$$

The continuous embedding $W^{1,p}(\Omega) \subset L^2(\Omega)$ for $p \geq \frac{2d}{d+2}$ implies

$$\|\overline{u} - u_h\|_2 \leq c \|\overline{u} - u_h\|_{1,p} \leq c h^{\min\{1, \frac{2}{p}\}}. \tag{7.59}$$

Note the following elementary inequality for all $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \boldsymbol{\eta} \in \mathbb{R}^d$:

$$\left||\boldsymbol{\xi}_1 - \boldsymbol{\eta}|^2 - |\boldsymbol{\xi}_2 - \boldsymbol{\eta}|^2\right| = \left|(\boldsymbol{\xi}_1 + \boldsymbol{\xi}_2 - 2\boldsymbol{\eta}) \cdot (\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2)\right| \leq 2(|\boldsymbol{\xi}_1| + |\boldsymbol{\xi}_2| + |\boldsymbol{\eta}|)|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2|.$$

From this and (7.59) we infer the estimate

$$\begin{aligned} \left|J(\overline{q}, \overline{u}) - J(\overline{q}, u_h)\right| &= \left|\frac{1}{2}\int_\Omega |\overline{u} - u_d|^2 \,\mathrm{d}x - \frac{1}{2}\int_\Omega |u_h - u_d|^2 \,\mathrm{d}x\right| \\ &\leq \int_\Omega (|\overline{u}| + |u_h| + |u_d|)|\overline{u} - u_h| \,\mathrm{d}x \\ &\leq (\|\overline{u}\|_2 + \|u_h\|_2 + \|u_d\|_2)\|\overline{u} - u_h\|_2 \\ &\leq c h^{\min\{1, \frac{2}{p}\}}, \end{aligned} \tag{7.60}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

where we have also used (7.17) and (7.29) for the last inequality. As the pair $(\overline{q}, u_h)$ is admissible for $(P_s)$, the inequality

$$J(\overline{q}_h, \overline{u}_h) \leq J(\overline{q}, u_h) \tag{7.61}$$

is fulfilled, and, consequently,

$$J(\overline{q}_h, \overline{u}_h) - J(\overline{q}, \overline{u}) \leq J(\overline{q}, u_h) - J(\overline{q}, \overline{u}) \overset{(7.60)}{\leq} ch^{\min\{1, \frac{2}{p}\}}. \tag{7.62}$$

Note that $\|\overline{q}_h\|_{\max\{p',2\}} \leq C$ uniformly in $h \in (0,1]$. To obtain the reverse inequality of (7.62), starting from $(\overline{q}_h, \overline{u}_h)$, we construct $(\overline{q}_h, \hat{u})$ by defining $\hat{u} \in \mathcal{V}$ as the solution to (7.16). Note that $(\overline{q}_h, \hat{u})$ are feasible for the exact optimal control problem $(P)$, although both $\overline{q}_h$ and $\hat{u}$ depend on $h$. As a result, we have

$$J(\overline{q}, \overline{u}) \leq J(\overline{q}_h, \hat{u}). \tag{7.63}$$

We can precisely use the same arguments as for (7.60) to obtain

$$\left|J(\overline{q}_h, \overline{u}_h) - J(\overline{q}_h, \hat{u})\right| \leq ch^{\min\{1, \frac{2}{p}\}}. \tag{7.64}$$

Combining inequalities (7.60), (7.61), (7.63), and (7.64), we finally arrive at

$$\begin{aligned} -ch^{\min\{1, \frac{2}{p}\}} &\overset{(7.60)}{\leq} J(\overline{q}, \overline{u}) - J(\overline{q}, u_h) \overset{(7.61)}{\leq} J(\overline{q}, \overline{u}) - J(\overline{q}_h, \overline{u}_h) \\ &\overset{(7.63)}{\leq} J(\overline{q}_h, \hat{u}) - J(\overline{q}_h, \overline{u}_h) \overset{(7.64)}{\leq} ch^{\min\{1, \frac{2}{p}\}}. \end{aligned}$$

This establishes the statement. □

Now let us deal with the case where the control space is discretized. To this end, we adapt the theory presented in [36] to our situation. To quantify the order of convergence, some regularity of the optimal control $\overline{q}$ is usually required. In the linear setting, additional regularity of $\overline{q}$ can be proven by deriving additional regularity of the adjoint state $z$. As we have seen in our discussion in Section 7.5, such additional regularity can only be shown in the case of the variational discretization.

**Theorem 7.7.3** (Convergence rates for piecewise constant controls). *For $\varepsilon \in [0, \infty)$ and $p \in [\frac{2d}{d+2}, \infty)$, let Assumptions 7.2.1 and 7.2.4 be satisfied. For each $h > 0$ and $\mathcal{Q}_{h,\mathrm{ad}} = \mathcal{Q}^0_{h,\mathrm{ad}}$, let $\overline{q}_h \in \mathcal{Q}_{h,\mathrm{ad}}$ be a discrete optimal control, and let $\overline{u}_h = u_h(\overline{q}_h) \in \mathcal{V}_h$ be the corresponding discrete optimal state, i. e., $(\overline{q}_h, \overline{u}_h)$ are global solutions of $(P_h)$. Further, let $(\overline{q}, \overline{u}) \in \mathcal{Q}_{\mathrm{ad}} \times \mathcal{V}$ be a global solution of $(P)$. Then there exists a constant $c > 0$ independent of $h$ such that*

$$\left|J(\overline{q}, \overline{u}) - J(\overline{q}_h, \overline{u}_h)\right| \leq ch^{\min\{1, \frac{1}{p-1}\}}. \tag{7.65}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*Proof.* We have already proven the existence of an accumulation point $(\overline{q}, \overline{u})$ in Theorem 7.7.1 and assume from now on that $(\overline{q}_h, \overline{u}_h)$ converges to this limit.

Let $(\hat{q}_h, \hat{u}_h) \in \mathcal{Q}_{h,\mathrm{ad}} \times \mathcal{V}$ be a global solution of the following auxiliary problem in which only the control variable is discretized:

$$\text{Minimize } J(q_h, u_h) \quad \text{subject to (7.16) and } (q_h, u_h) \in \mathcal{Q}_{h,\mathrm{ad}} \times \mathcal{V}. \tag{7.66}$$

To derive the stated error estimate, we split the error as follows:

$$\left|J(\overline{q}, \overline{u}) - J(\overline{q}_h, \overline{u}_h)\right| \leq \left|J(\overline{q}, \overline{u}) - J(\hat{q}_h, \hat{u}_h)\right| + \left|J(\hat{q}_h, \hat{u}_h) - J(\overline{q}_h, \overline{u}_h)\right|. \tag{7.67}$$

By repeating the proof of Theorem 7.7.2 we can estimate the second term on the right-hand side of (7.67) as

$$\left|J(\hat{q}_h, \hat{u}_h) - J(\overline{q}_h, \overline{u}_h)\right| \leq ch^{\min\{1, \frac{2}{p}\}}. \tag{7.68}$$

This is possible as all constants appearing in this proof are only dependent of $\|\hat{q}_h\|_{\max\{p',2\}}$ (and the regularity of $\mathbb{T}_h$ and characteristics of $\boldsymbol{S}$). Note that $\|\hat{q}_h\|_{\max\{p',2\}}$ is uniformly bounded in $h \in (0, 1]$.

Thus it is sufficient to estimate the first term on the right-hand side of (7.67). To this end, we use again similar arguments as in the proof of Theorem 7.7.2. Let us set $q_h := \Pi_h \overline{q}$, where $\Pi_h$ stands for the $L^2$-projection onto $Q_h^0$. It is clear that $\Pi_h \colon \mathcal{Q}_{\mathrm{ad}} \to \mathcal{Q}_{h,\mathrm{ad}}^0$.

Let $u_h \in \mathcal{V}$ be the solution to the state equation (7.16) for control $q_h$. From Lemma 7.3.2 we deduce the estimate

$$\|\nabla \overline{u} - \nabla u_h\|_p \lesssim \begin{cases} \|\varepsilon + |\nabla \overline{u}| + |\nabla u_h|\|_p^{2-p} \|\overline{q} - \Pi_h \overline{q}\|_{-1,p'} & \text{for } p \leq 2, \\ \|\overline{q} - \Pi_h \overline{q}\|_{-1,p'}^{\frac{1}{p-1}} & \text{for } p \geq 2. \end{cases} \tag{7.69}$$

Due to the uniform a priori bounds (7.17) and (7.29) and the stability of $\Pi_h$, in the case $p \leq 2$, there exists a constant $C > 0$ independent of $h$ such that

$$\left\|\varepsilon + |\nabla \overline{u}| + |\nabla u_h|\right\|_p^{2-p} \leq C.$$

Employing Lemma 7.4.2, we can bound the right-hand side of (7.69) by

$$\|\nabla \overline{u} - \nabla u_h\|_p \leq ch^{\min\{1, \frac{1}{p-1}\}}. \tag{7.70}$$

From the convexity of $J$ we conclude

$$\begin{aligned} J(\overline{q}, \overline{u}) &\geq J(q_h, u_h) + \left\langle J'(q_h, u_h), (\overline{q} - q_h, \overline{u} - u_h) \right\rangle \\ &= J(q_h, u_h) + \alpha(q_h, \overline{q} - q_h)_\Omega + (u_h - u_d, \overline{u} - u_h)_\Omega. \end{aligned}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Since $(q_h, u_h)$ is feasible for (7.66), the inequality

$$\begin{aligned} 0 \leq J(\hat{q}_h, \hat{u}_h) - J(\overline{q}, \overline{u}) &\leq J(q_h, u_h) - J(\overline{q}, \overline{u}) \\ &\leq \alpha(q_h, q_h - \overline{q})_\Omega + (u_h - u_d, u_h - \overline{u})_\Omega \end{aligned}$$

follows. The last term on the right-hand side is bounded by

$$(u_h - u_d, u_h - \overline{u})_\Omega \leq \|u_h - u_d\|_2 \|u_h - \overline{u}\|_2 \overset{(7.70)}{\leq} ch^{\min\{1, \frac{1}{p-1}\}}$$

for $p \geq \frac{2d}{d+2}$. Moreover,

$$(q_h, q_h - \overline{q})_\Omega = (q_h, \Pi_h \overline{q} - \overline{q})_\Omega = 0$$

due to the definition of the $L^2$-projection. To sum up, we get

$$0 \leq J(\hat{q}_h, \hat{u}_h) - J(\overline{q}, \overline{u}) \leq ch^{\min\{1, \frac{1}{p-1}\}}. \tag{7.71}$$

Combining (7.67), (7.68), and (7.71), we conclude the statement noting that $h^{\min\{1, \frac{1}{p-1}\}} \geq h^{\min\{1, \frac{2}{p}\}}$. □

# Bibliography


[1] J. W. Barrett and W. B. Liu. Quasi-norm error bounds for the finite element approximation of a non-Newtonian flow. *Numer. Math.*, 68:437–456, 1994.

[2] L. Belenki, L. C. Berselli, L. Diening, and M. Růžička. On the finite element approximation of $p$-Stokes systems. *SIAM J. Numer. Anal.*, (2):373–397, 2012.

[3] L. Belenki, L. Diening, and C. Kreuzer. Optimality of an adaptive finite element method for the $p$-Laplacian equation. *IMA J. Numer. Anal.*, 32(2): 484–510, 2012.

[4] L. C. Berselli, L. Diening, and M. Růžička. Existence of strong solutions for incompressible fluids with shear dependent viscosities. *J. Math. Fluid Mech.*, 12:101–132, 2010.

[5] T. Bhattacharya and F. Leonetti. A new Poincaré inequality and its application to the regularity of minimizers of integral functionals with nonstandard growth. *Nonlinear Anal.*, 17:833–839, 1991.

[6] L. Bonifacius and I. Neitzel. Second order optimality conditions for optimal control of quasilinear parabolic equations. *Math. Control Relat. Fields*, 8(1):1–34, 2018.

[7] S. Brenner and R. L. Scott. *The Mathematical Theory of Finite Element Methods*. Springer Verlag, Berlin, Heidelberg, New York, 1994.

[8] E. Casas and K. Chrysafinos. Analysis and optimal control of some quasilinear parabolic equations. *Math. Control Relat. Fields*, 8(3–4):607–623, 2018.

[9] E. Casas and V. Dhamo. Optimality conditions for a class of optimal boundary control problems with quasilinear elliptic equations. *Control Cybern.*, 40(2):457–490, 2011.

[10] E. Casas, P. I. Kogut, and G. Leugering. Approximation of optimal control problems in the coefficient for the $p$-Laplace equation. I. Convergence result. *SIAM J. Control Optim.*, 54(3):1406–1422, 2016.


EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[11] E. Casas and F. Tröltzsch. First- and second-order optimality conditions for a class of optimal control problems with quasilinear elliptic equations. *SIAM J. Control Optim.*, 48(2):688–718, 2009.
[12] S.-S. Chow. Finite element error estimates for non-linear elliptic equations of monotone type. *Numer. Math.*, 54:373–393, 1989.
[13] A. Cianchi and V. Maz'ya. Second-order two-sided estimates in nonlinear elliptic problems. *Arch. Ration. Mech. Anal.*, 229:569–599, 2018.
[14] A. Cianchi and V. G. Maz'ya. Global Lipschitz regularity for a class of quasilinear elliptic equations. *Commun. Partial Differ. Equ.*, 36(1):100–133, 2010.
[15] P. G. Ciarlet. *The finite elements methods for elliptic problems*. North-Holland, 1980.
[16] F. Crispo, C. R. Grisanti, and P. Maremonti. On the high regularity of solutions to the $p$-Laplacian boundary value problem in exterior domains. *Ann. Mat. Pura Appl.*, 195:821–834, 2016.
[17] J. C. De los Reyes. *Numerical PDE-constrained optimization*. Springer International Publishing, 2015.
[18] K. Deckelnick and M. Hinze. Convergence of a finite element approximation to a state-constrained elliptic control problem. *SIAM J. Numer. Anal.*, 45(5):1937–1953, 2007.
[19] L. Diening, C. Ebmeyer, and M. Růžička. Optimal convergence for the implicit space-time discretization of parabolic systems with $p$-structure. *SIAM J. Numer. Anal.*, 45(2):457–472, 2007.
[20] L. Diening and F. Ettwein. Fractional estimates for non-differentiable elliptic systems with general growth. *Forum Math.*, 20:523–556, 2008.
[21] L. Diening, D. Kröner, M. Růžička, and I. Toulopoulos. A local discontinuous Galerkin approximation for systems with $p$-structure. *IMA J. Numer. Anal.*, 34(4):1447–1488, 2014.
[22] L. Diening and M. Růžička. Interpolation operators in Orlicz–Sobolev spaces. *Numer. Math.*, 107(1):107–129, 2007.
[23] L. Diening, B. Stroffolini, and A. Verde. Everywhere regularity of functionals with $\varphi$-growth. *Manuscr. Math.*, 129:449–481, 2009.
[24] C. Ebmeyer and W. B. Liu. Quasi-norm interpolation error estimates for piecewise linear finite element approximation of $p$-Laplacian problems. *Numer. Math.*, 100(204):233–258, 2005.
[25] B. Endtmayer, U. Langer, I. Neitzel, T. Wick, and W. Wollner. Multigoal-oriented optimal control problems with nonlinear PDE constraints. *Comput. Math. Appl.*, 79(10):3001–3026, 2020.
[26] C. Helanow and J. Ahlkrona. Stabilized equal low-order finite elements in ice sheet modeling – accuracy and robustness. *Comput. Geosci..*, 22:951–974, 2018.
[27] R. Herzog and W. Wollner. A conjugate direction method for linear systems in Banach spaces. *J. Inverse Ill-Posed Probl.*, 25(5):553–572, 2017.
[28] M. Hinze. A variational discretization concept in control constrained optimization: the linear-quadratic case. *Comput. Optim. Appl.*, 30(1):45–61, 2005.
[29] A. Hirn. Approximation of the $p$-Stokes equations with equal-order finite elements. *J. Math. Fluid Mech.*, 15:65–88, 2013.
[30] A. Hirn. Finite element approximation of singular power-law systems. *Math. Comput.*, 82:1247–1268, 2013.
[31] J. L. Lions. *Optimal Control of Systems Governed by Partial Differential Equations*. Die Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 1971.
[32] W. B. Liu. Degenerate quasilinear elliptic equations arising from bimaterial problems in elastic–plastic mechanics. *Nonlinear Anal.*, 35:517–529, 1999.
[33] W. B. Liu and J. W. Barrett. A remark on the regularity of the solutions of the $p$-Laplacian and its applications to their finite element approximation. *J. Math. Anal. Appl.*, 178(2):470–487, 1993.
[34] J. Málek, J. Nečas, M. Rokyta, and M. Růžička. *Weak and measure-valued solutions to evolutionary PDEs*, volume 13 of Applied Mathematics and Mathematical Computation. Chapman and Hall, London, 1996.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[35] C. Meyer. Error estimates for the finite-element approximation of an elliptic control problem with pointwise state and control constraints. *Control Cybern.*, 37(1):51–85, 2008.
[36] C. Ortner and W. Wollner. A priori error estimates for optimal control problems with pointwise constraints on the gradient of the state. *Numer. Math.*, 118:587–600, 2011.
[37] R. Rannacher. *Numerik 2. Numerik partieller Differentialgleichungen*. Lecture Notes Mathematik. Heidelberg University Publishing, 2017.
[38] M. Růžička. *Nichtlineare Funktionalanalysis*. Springer-Verlag, 2004.
[39] L. R. Scott and S. Zhang. Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Math. Comput.*, 54(190):483–493, 1990.
[40] F. Tröltzsch. *Optimal Control of Partial Differential Equations*, volume 112 of Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2010.
[41] E. Zeidler. *Nonlinear Functional Analysis and its Applications. II/B: Nonlinear Monotone Operators*. Springer-Verlag, New York, 1990.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Ulrich Langer, Olaf Steinbach, Fredi Tröltzsch, and Huidong Yang

# 8 Unstructured space-time finite element methods for optimal sparse control of parabolic equations

**Abstract:** We consider a space-time finite element method on fully unstructured simplicial meshes for optimal sparse control of semilinear parabolic equations. The objective is a combination of a standard quadratic tracking-type functional including a Tikhonov regularization term and of the $L^1$-norm of the control that accounts for its spatio-temporal sparsity. We use a space-time Petrov–Galerkin finite element discretization for the first-order necessary optimality system of the optimal sparse control problem. The discretization is based on a variational formulation that employs continuous piecewise linear finite elements simultaneously in space and time. Finally, we solve the discretized nonlinear optimality system that consists of coupled forward–backward state and adjoint state equations by a semismooth Newton method.

**Keywords:** space-time finite element method, optimal sparse control, semilinear parabolic equations

**MSC 2010:** 49J20, 35K20, 65M60, 65M50, 65M15, 65Y05

## 8.1 Introduction

Optimal sparse control of partial differential equations with the $L^1$-norm of the control in the objective functional was first analyzed in [31], about a decade ago. That paper, where a linear elliptic state equation is considered, initiated an active research on sparsity in the control of PDEs. The method was extended to semilinear elliptic optimal control problems in [10] and to problems governed by elliptic equations with un-

**Acknowledgement:** We greatly acknowledge the support from Johann Radon Institute for Computational and Applied Mathematics (RICAM) during the special semester on optimization that took place between 14th October and 11th December, 2019, at RICAM in Linz, Austria. The authors thank the anonymous reviewers for careful reading and many helpful comments, which helped us to improve the paper considerably.

**Ulrich Langer, Huidong Yang,** RICAM, Austrian Academy of Sciences, Altenberger Straße 69, 4040 Linz, Austria, e-mails: ulrich.langer@ricam.oeaw.ac.at, huidong.yang@ricam.oeaw.ac.at
**Olaf Steinbach,** Institut für Angewandte Mathematik, Technische Universität Graz, Steyrergasse 30, 8010 Graz, Austria, e-mail: o.steinbach@tugraz.at
**Fredi Tröltzsch,** Institut für Mathematik, Technische Universität Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany, e-mail: troeltzsch@math.tu-berlin.de

https://doi.org/10.1515/9783110695984-008

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

certain coefficients in [28]. In [15, 16, 30], problems of optimal sparse control were investigated for the Schlögl and FitzHugh–Nagumo systems, where traveling wave fronts or spiral waves were controlled. An advantage of sparsity is that the optimal control concentrates on regions where it can have the highest impact for optimality, whereas it is exactly zero in the other parts. In applications, sparsity restricts the set of points where the control has to act. This may be useful for the technical implementation of the control.

In this paper, we concentrate on spatio-temporal sparsity promoted by including the $L^1$-norm of the control in the objective functional. This type of sparsity fits best to our numerical method motivated by equations of mathematical physics that exhibit spatio-temporally moving objects as solutions. Another type of sparsity is the directional sparsity, which has been investigated, e. g., in [11, 13, 14, 21]. We briefly discuss this issue in Section 8.2. We also mention another class of sparse optimal control approaches with control in measure spaces; see, e. g., [5, 8, 12, 17, 24]. A thorough review of the existing literature on this challenging topic is beyond the scope of this work. Therefore, we refer to the recent survey [7] and the references therein on sparse solutions in optimal control of both elliptic and parabolic equations.

Moreover, numerical approximations of optimal sparse controls of elliptic and parabolic problems were of great interest. For example, the standard five-point stencil was used in [31] for the discretization of elliptic optimal sparse control problems. In [10], rigorous error estimates were proved for the finite element approximation of semilinear elliptic sparse control problems with box constraints on the control. In the discretized optimality conditions, continuous piecewise linear approximations are used for the state and adjoint state, whereas a piecewise constant ansatz was applied to the control and subdifferential. We also mention the approximation of sparse controls by piecewise linear functions in [9]. Here special quadrature formulae are adopted to discretize the squared $L^2$-norm and the $L^1$-norm of the control in the objective functional. This leads to an elementwise representation of the control and subdifferential.

Later, error estimates were derived for the space-time finite element approximation of parabolic optimal sparse control problems without control constraints in [13]. The discretization was performed on tensor-structured space-time meshes. In the associated discretized optimal control problem, there was used a space-time finite element ansatz for the state that consists of products of continuous piecewise linear basis functions in space and piecewise constant basis functions in time. For the control, piecewise constant approximations in both spatial and temporal directions were utilized. The same space as for the state was employed for the adjoint state in the discretized optimality system. Improved approximation rates were achieved in [14]. For the control discretization, basis functions that are continuous and piecewise linear in space and piecewise constant in time were employed. These methods can be reinterpreted as an implicit Euler discretization of the spatially discretized optimality system.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

For optimal sparse control of the Schlögl and FitzHugh–Nagumo models considered in [15], a semiimplicit Euler-method in time and continuous piecewise linear finite elements in space were applied to both the state and adjoint state equations. Recently, in [40], for the optimal control of the convective FitzHugh–Nagumo equations, the state and adjoint state equations were discretized by a symmetric interior penalty Galerkin method in space and the backward Euler method in time.

In contrast to the discretization methods discussed, we apply continuous space-time finite element approximations on fully unstructured simplicial meshes for parabolic optimal sparse control problems with control constraints. This can be seen as an extension of the Petrov–Galerkin space-time finite element method proposed in [33] for parabolic problems and in our recent work [26] for parabolic optimal control problems. This kind of unstructured space-time finite element approaches has gained increasing interest; see, e. g., [2, 3, 23, 25, 35, 37, 41] and the survey [36].

In comparison to the more conventional time-stepping methods or tensor-structured space-time methods [18, 19, 29], this unstructured space-time approach provides us with more flexibility in constructing parallel space-time solvers such as parallel space-time algebraic multigrid preconditioners [25] or space-time balancing domain decomposition by constraints (BDDC) preconditioners [27]. Moreover, it becomes more convenient to realize simultaneous space-time adaptivity on unstructured space-time meshes [25, 26, 34] than the other methods. In fact, time is just considered as another spatial coordinate. For more comparisons of our space-time finite element methods with others, we refer to [36].

The remainder of this paper is structured as follows. Section 8.2 describes a model optimal sparse control problem that we aim to solve. Some preliminary existing results concerning optimality conditions are given in Section 8.3. The space-time finite element discretization of the associated optimality system, the discretized optimality conditions, and the application of the semismooth Newton iteration are discussed in Section 8.4. The applicability of our proposed method is confirmed by two numerical examples in Section 8.5. Finally, some conclusions are drawn in Section 8.6.

## 8.2 The optimal sparse control model problem

We consider the optimal sparse control problem

$$\min_{z\in Z_{\mathrm{ad}}} \mathcal{J}(z) := \frac{1}{2}\,\|u_z - u_Q\|^2_{L^2(Q)} + \frac{\rho}{2}\,\|z\|^2_{L^2(Q)} + \mu\,\|z\|_{L^1(Q)}, \tag{8.1}$$

where the admissible set of controls is

$$Z_{\mathrm{ad}} = \{z \in L^\infty(Q) : a \le z(x,t) \le b \text{ for almost all } (x,t) \in Q\}, \tag{8.2}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

and $u_z$ is the unique solution of the semilinear state equation

$$\begin{aligned} \partial_t u - \Delta_x u + R(u) &= z \quad \text{in } Q := \Omega \times (0,T), \\ u &= 0 \quad \text{on } \Sigma := \partial\Omega \times (0,T), \\ u &= u_0 \quad \text{on } \Sigma_0 := \Omega \times \{0\}. \end{aligned} \tag{8.3}$$

Here the spatial computational domain $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,3\}$, is supposed to be bounded and Lipschitz, $T > 0$ is the fixed terminal time, $\partial_t$ denotes the partial time derivative, $\Delta_x = \sum_{i=1}^d \partial_{x_i}^2$ is the spatial Laplacian, and the distributed control $z$ acts as a source term in $Q$. Moreover, $u_Q \in L^2(Q)$ is a given desired state. We further assume that

$$-\infty < a < 0 < b < +\infty, \quad \rho > 0, \quad \mu > 0.$$

The nonlinear reaction term $R$ is defined by

$$R(u) = (u - u_1)(u - u_2)(u - u_3)$$

with given real numbers $u_1 \le u_2 \le u_3$. The term $\mu\|z\|_{L^1(Q)}$ in the objective functional accounts for the spatio-temporal sparsity of optimal controls. Notice that the functional $g : L^1(Q) \to \mathbb{R}$ defined by $g(\cdot) = \|\cdot\|_{L^1(Q)}$ is Lipschitz continuous and convex but not Fréchet differentiable. Similar model problems for semilinear equations have been studied, e. g., in [7, 11].

In this setting, $\rho$ plays the role of a regularization parameter, whereas $\mu$ accounts for sparsity. Often, $\rho$ is considered as the cost of the control. From a mathematical point of view, this parameter stands for higher regularity of optimal controls and for the stability of numerical solution algorithms. A similar interpretation of the sparsity parameter $\mu$ is difficult; it may be viewed as a price for the size of the region where the control is allowed to act. The larger $\mu$ is, the larger is the set of points where optimal controls vanish. An inspection of relation (8.6a) below reveals that optimal controls are exactly zero in all the points where the absolute value of the associated adjoint state is not greater than $\mu$. On the other hand, an increase of $\mu$ worsens the approximation of the desired target state $u_Q$, because the control is restricted to a smaller set and hence less flexible. The selection of a good sparsity parameter $\mu$ depends on the specific application and can be found by numerical tests.

We concentrate on spatio-temporal sparsity, because our focus is on completely unstructured spatio-temporal grids. Moreover, this fits best to problems with spatio-temporally moving targets or state functions. The control of such moving objects motivated us to our research. They occur in some reaction–diffusion equations, for instance, in the Schlögl model or the FitzHugh–Nagumo equations of mathematical physics; see, e. g., [15, 16, 26, 30, 40]. These equations and associated control problems have applications in chemistry, physics, physiology, and medicine. Sparsity is helpful, because it can lead to a better technical implementation of controls.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

The so-called directional sparsity leads to another form of sparse controls. For instance, temporal sparsity is generated by the functional

$$g(z) = \int_0^T \left( \int_\Omega |z(x,t)|^2 \, dx \right)^{1/2} dt$$

defined on $L^1(0,T;L^2(\Omega))$, whereas spatial sparsity is obtained by the choice

$$g(z) = \int_\Omega \left( \int_0^T |z(x,t)|^2 \, dt \right)^{1/2} dx$$

in $L^1(\Omega;L^2(0,T))$. The directional sparsity was introduced by Herzog et al. [21] and further investigated in [7, 14, 20, 24].

## 8.3 Preliminary results

Let us recall some facts from the literature, e. g., [7]. For all $z \in L^p(Q)$, $p > d/2 + 1$, the state equation (8.3) has a unique solution $u_z \in W(0,T) \cap L^\infty(Q)$, where

$$W(0,T) = \{v \in L^2(0,T;H_0^1(\Omega)) : \partial_t v \in L^2(0,T;H^{-1}(\Omega))\}. \tag{8.4}$$

The mapping $z \mapsto u_z$ is continuously Fréchet differentiable in these spaces. The optimal control problem has at least one (globally) optimal control denoted by $\bar{z}$, and the associated optimal state is denoted by $\bar{u}$.

If $\bar{z}$ is a locally optimal control of the model problem, then there exist a unique adjoint state $\bar{p} \in W(0,T)$ and $\bar{\lambda} \in \partial g(\bar{u}) \subset L^\infty(Q)$ such that $(\bar{u},\bar{p},\bar{z},\bar{\lambda})$ solves the optimality system

$$\partial_t u - \Delta_x u + R(u) = z \quad \text{in } Q, \quad u = 0 \quad \text{on } \Sigma, \quad u = u_0 \quad \text{on } \Sigma_0, \tag{8.5a}$$

$$-\partial_t p - \Delta_x p + R'(u)p = u - u_Q \quad \text{in } Q, \quad p = 0 \quad \text{on } \Sigma, \quad p = 0 \quad \text{on } \Sigma_T, \tag{8.5b}$$

$$\int_Q (p + \rho z + \mu\lambda)(v - z) \, dx \, dt \geq 0 \quad \text{for all } v \in Z_{\text{ad}}, \tag{8.5c}$$

where $\Sigma_T := \Omega \times \{T\}$. A detailed discussion of this optimality system leads to the relations

$$\bar{z}(x,t) = 0 \Leftrightarrow |\bar{p}(x,t)| \leq \mu, \tag{8.6a}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$$\bar{z}(x,t) = \mathbf{Proj}_{[a,b]}\left(-\frac{1}{\rho}(\bar{p}(x,t) + \mu\bar{\lambda}(x,t))\right), \tag{8.6b}$$

$$\bar{\lambda}(x,t) = \mathbf{Proj}_{[-1,1]}\left(-\frac{1}{\mu}\bar{p}(x,t)\right), \tag{8.6c}$$

which hold for almost all $(x,t) \in Q$; see, e. g., [15]. Here the projection operator $\mathbf{Proj}_{[\alpha,\beta]} : \mathbb{R} \to [\alpha,\beta]$ is defined by $\mathbf{Proj}_{[\alpha,\beta]}(q) = \max\{\alpha, \min\{q,\beta\}\}$; see, e. g., [38]. The subdifferential $\partial g(z)$ of the $L^1$-norm at the control $z \in L^1(Q)$ belongs to $L^\infty(Q)$ and is given as follows:

$$\lambda \in \partial g(z) \Leftrightarrow \begin{cases} \lambda(x,t) = 1 & \text{if } z(x,t) > 0, \\ \lambda(x,t) \in [-1,1] & \text{if } z(x,t) = 0, \\ \lambda(x,t) = -1 & \text{if } z(x,t) < 0 \end{cases}$$

for almost all $(x,t) \in Q$. By these relations, we obtain the following form of an optimal control:

$$\bar{z} = \begin{cases} b & \text{on } \mathcal{A}_b := \{(x,t) \in Q : \bar{p}(x,t) < -\rho b - \mu\}, \\ -\frac{1}{\rho}(\bar{p} + \mu) & \text{on } \mathcal{I}_+ := \{(x,t) \in Q : -\rho b - \mu \le \bar{p}(x,t) < -\mu\}, \\ 0 & \text{on } \mathcal{A}_0 := \{(x,t) \in Q : |\bar{p}(x,t)| \le \mu\}, \\ -\frac{1}{\rho}(\bar{p} - \mu) & \text{on } \mathcal{I}_- := \{(x,t) \in Q : \mu < \bar{p}(x,t) \le -\rho a + \mu\}, \\ a & \text{on } \mathcal{A}_a := \{(x,t) \in Q : \bar{p}(x,t) > -\rho a + \mu\}. \end{cases} \tag{8.7}$$

The set $\mathcal{A}_0$ accounts for the sparsity of the control.

For convenience, we define the real function $\mathbb{P} : \mathbb{R} \to [a,b]$ by

$$\mathbb{P}(s) = \mathbf{Proj}_{[a,b]}\left(-\frac{1}{\rho}\left(s + \mu\mathbf{Proj}_{[-1,1]}\left(-\frac{1}{\mu}s\right)\right)\right).$$

Eliminating the control from the optimality system and using the above projection formulae, we obtain the following system for the state and the adjoint state:

$$\partial_t u - \Delta_x u + R(u) = \mathbb{P}(p) \quad \text{in } Q, \tag{8.8a}$$

$$u = 0 \quad \text{on } \Sigma, \quad u = u_0 \quad \text{on } \Sigma_0,$$

$$-\partial_t p - \Delta_x p + R'(u)p = u - u_Q \quad \text{in } Q, \tag{8.8b}$$

$$p = 0 \quad \text{on } \Sigma, \quad p = 0 \quad \text{on } \Sigma_T.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Let us define the Bochner spaces for the state and adjoint state variables as follows:

$$\begin{aligned} X_0 &:= L^2(0,T;H_0^1(\Omega)) \cap H_{0,}^1(0,T;H^{-1}(\Omega)) = \{v \in W(0,T), v = 0 \text{ on } \Sigma_0\},\\ X_T &:= L^2(0,T;H_0^1(\Omega)) \cap H_{,0}^1(0,T;H^{-1}(\Omega)) = \{v \in W(0,T), v = 0 \text{ on } \Sigma_T\},\\ Y &:= L^2(0,T;H_0^1(\Omega)). \end{aligned}$$

The space-time variational formulation for the coupled system (8.8) reads: Find $u \in X_0$ and $p \in X_T$ such that the variational equations

$$\int_Q \partial_t u\, v\, dx\, dt + \int_Q \nabla_x u \cdot \nabla_x v\, dx\, dt + \int_Q R(u)\, v\, dx\, dt = \int_Q \mathbb{P}(p)\, v\, dx\, dt, \tag{8.9a}$$

$$\begin{aligned} &-\int_Q u\, q\, dx\, dt - \int_Q \partial_t p\, q\, dx\, dt + \int_Q \nabla_x p \cdot \nabla_x q\, dx\, dt + \int_Q R'(u) p\, q\, dx\, dt \\ &\quad = -\int_Q u_Q\, q\, dx\, dt \end{aligned} \tag{8.9b}$$

hold for all $v, q \in Y$. This system is solvable, because the optimal control problem has at least one solution.

Preparing the discussion of the semismooth Newton method for solving the associated discretized system, we briefly explain this method for the continuous setting. First, we take a look on the structure of $\mathbb{P}$. By a simple case study we find

$$\mathbb{P}(s) = \begin{cases} b, & s < -\rho b - \mu, \\ -\frac{s+\mu}{\rho}, & -\rho b - \mu \le s < -\mu, \\ 0, & -\mu \le s \le \mu, \\ -\frac{s-\mu}{\rho}, & \mu < s \le \mu - \rho a, \\ a, & \mu - \rho a < s. \end{cases}$$

Therefore, $\mathbb{P}$ is a piecewise linear and continuous function that is continuously differentiable in $\mathbb{R} \setminus \{-\rho b - \mu, -\mu, \mu, \mu - \rho a\}$. Between the corners of $\mathbb{P}$, the derivative $\mathbb{P}'(s)$ is equal to 0, $-1/\rho$, 0, $-1/\rho$, 0, in the same order as above, as it is obvious from the representation of $\mathbb{P}$.

If $\mathbb{P}$ were a differentiable function, then the classical Newton iteration for solving system (8.8) would read as follows: Given the last iterate $(u_n, p_n)$, the next iterate $(u, p) = (u_{n+1}, p_{n+1})$ would be obtained as the unique solution of

$$\begin{aligned} \partial_t u - \Delta_x u + R(u_n) + R'(u_n)(u - u_n) &= \mathbb{P}(p_n) + \mathbb{P}'(p_n)(p - p_n), \\ -\partial_t p - \Delta_x p + R'(u_n)p + R''(u_n)p_n(u - u_n) &= u - u_Q \end{aligned} \tag{8.10}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

in $Q$ subject to $u(0) = u_0$, $p(T) = 0$, and homogeneous Dirichlet boundary conditions. As a piecewise linear function, $\mathbb{P}$ is not differentiable in the corner points mentioned above, but it is semismooth; cf. [22, §18.2] or [39, §2.2]. The associated generalized derivative is set-valued in the corners. We are justified to select in particular the values of the directional derivatives of $\mathbb{P}$ in the corner points and define formally

$$\mathbb{P}'(s) := \begin{cases} -\frac{1}{\rho}, & s = -\rho b - \mu, \\ 0, & s = -\mu, \\ 0, & s = \mu, \\ -\frac{1}{\rho}, & s = \mu - \rho a. \end{cases}$$

Inserting this definition of $\mathbb{P}'$ into the above Newton iteration scheme, we arrive at the semismooth Newton method described in detail in [22] and [39]. Moreover, evaluating $\mathbb{P}(u_n) + \mathbb{P}'(u_n)(u - u_n)$ in the right-hand side of (8.10) reveals that the semismooth Newton iteration is equivalent to an active set method. For its presentation, we introduce the following sets depending on the iteration number $n$:

$$\begin{cases} \mathcal{A}_a^n := \{(x,t) \in Q : \ p_n(x,t) > \mu - \rho a\}, \\ \mathcal{A}_b^n := \{(x,t) \in Q : \ p_n(x,t) < -\mu - \rho b\}, \\ \mathcal{A}_0^n := \{(x,t) \in Q : \ |p_n(x,t)| \le \mu\}, \\ \mathcal{I}_-^n := \{(x,t) \in Q : \ \mu < p_n(x,t) \le \mu - \rho a\}, \\ \mathcal{I}_+^n := \{(x,t) \in Q : \ -\rho b - \mu \le p_n(x,t) < -\mu)\}. \end{cases}$$

Having the current iterate $(u_n, p_n)$, the next step of the active set strategy reads as follows: The new iterate $(u_{n+1}, p_{n+1})$ is the solution $(u, p)$ of the system

$$\begin{aligned} \partial_t u - \Delta_x u + R(u_n) + R'(u_n)(u - u_n) &= \begin{cases} b & \text{on } \mathcal{A}_b^n, \\ -\frac{1}{\rho}(p + \mu) & \text{on } \mathcal{I}_+^n, \\ 0 & \text{on } \mathcal{A}_0^n, \\ -\frac{1}{\rho}(p - \mu) & \text{on } \mathcal{I}_-^n, \\ a & \text{on } \mathcal{A}_a^n, \end{cases} \\ -\partial_t p - \Delta_x p + R'(u_n)p + R''(u_n)p_n(u - u_n) &= u - u_Q, \end{aligned} \tag{8.11}$$

endowed with the associated initial and boundary conditions mentioned above.

Due to the projection formula on the right-hand side of (8.9a), we need special care to discretize the optimality system. In fact, following the discretization scheme proposed in [1], we go back to the variational inequality (8.5c) and use a piecewise constant approximation for the control to derive the discretization of the first-order optimality system. We will discuss this in the forthcoming section.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

## 8.4 Space-time finite element discretization

For the space-time finite element discretization of the optimality system (8.8), we consider an admissible triangulation $\mathcal{T}_h(Q)$ of the space-time domain $Q$ into shape regular simplicial finite elements $\tau$. Here the mesh size $h$ is defined by $h = \max_{\tau\in\mathcal{T}_h} h_\tau$ with $h_\tau$ being the diameter of the element $\tau$; see, e. g., [6, 32]. For simplicity, we assume $\Omega$ to be a polygonal spatial domain. Therefore, the triangulation exactly covers $Q = \Omega \times (0, T)$.

Let $S_h^1(Q)$ be the space of continuous and piecewise linear functions that are defined with respect to the triangulation $\mathcal{T}_h(Q)$. The discretized variational form of the state equation (8.3) reads as follows: Find $u_h \in X_{0,h} = S_h^1(Q) \cap X_0$ such that

$$\int_Q \partial_t u_h\, v_h \,dx\,dt + \int_Q \nabla_x u_h \cdot \nabla_x v_h \,dx\,dt + \int_Q R(u_h)\, v_h \,dx\,dt = \int_Q z\, v_h \,dx\,dt \tag{8.12}$$

for all $v_h \in X_{0,h}$. For approximating the control $z$, we define the space

$$Z_h = \{z_h \in L^\infty(Q) : z_h \text{ is constant in each } \tau \in \mathcal{T}_h\}.$$

An element $z_h \in Z_h$ can be represented in the form

$$z_h = \sum_{\tau\in\mathcal{T}_h} \xi_\tau \mathcal{X}_\tau$$

with $\xi_\tau \in \mathbb{R}$ and $\mathcal{X}_\tau$ being the characteristic function of $\tau$. Moreover, the set of discretized admissible controls is defined by

$$Z_{\mathrm{ad},h} = \{z_h \in Z_h : a \le z_h|_\tau \le b \ \text{ for all } \tau \in \mathcal{T}_h\}.$$

The choice of piecewise constant admissible controls has the advantage that the optimality conditions can be formulated in a cellwise way. In particular, the discretized variational inequality (8.13c) below can be written in the cellwise form (8.14). This cellwise structure also holds for the projection formula (8.15) and for the form of the discretized subdifferential. Although piecewise linear controls fit better to a given finite element discretization, the associated optimality conditions and their implementation are more difficult, because the projection $\mathbb{P}$ would require cuts through the cells. For the approximation of sparse optimal controls by piecewise linear functions, we refer to [9].

Now we consider the discretized optimality system

$$\begin{aligned}\int_Q \partial_t u_h\, v_h \,dx\,dt + \int_Q \nabla_x u_h \cdot \nabla_x v_h \,dx\,dt + \int_Q R(u_h)\, v_h \,dx\,dt \\ = \int_Q z_h\, v_h \,dx\,dt \quad \text{for all } v_h \in X_{0,h},\end{aligned} \tag{8.13a}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$$-\int_Q u_h\, q_h\, dx\, dt - \int_Q \partial_t p_h\, q_h\, dx\, dt + \int_Q \nabla_x p_h \cdot \nabla_x q_h\, dx\, dt$$
$$+ \int_Q R'(u_h) p_h\, q_h\, dx\, dt = -\int_Q u_Q\, q_h\, dx\, dt \quad \text{for all } q_h \in X_{T,h}, \tag{8.13b}$$

$$\int_Q (p_h + \rho z_h + \mu\lambda_h)(v_h - z_h)\, dx\, dt \;\geq\; 0 \quad \text{for all } v_h \in Z_{\mathrm{ad},h} \tag{8.13c}$$

for an approximation of a locally optimal reference solution $(\bar{u}, \bar{p}, \bar{z}, \bar{\lambda})$ of the continuous optimality system. We tacitly assume the existence of such a discretized solution that is locally unique. This question of existence and local uniqueness will not be discussed in this paper.

Note that the discretized control $z_h$ can be equivalently expressed by the vector $\xi = (\xi_\tau)_{\tau\in\mathcal{T}_h}$. The $L^1$-norm of $z_h$ takes the form

$$g(z_h) = g_h(\xi) = \sum_{\tau\in\mathcal{T}_h} |\tau||\xi_\tau|$$

with $|\tau|$ being the area of a finite element $\tau$. Therefore, the subdifferential of $g_h(\xi)$ can be used in the subdifferential calculus for the discretized $L^1$-norm, and we need only a subset of elements of $\lambda_h \in \partial g(z_h)$ of the form

$$\lambda_h = \sum_{\tau\in\mathcal{T}_h} v_\tau \mathcal{X}_\tau \quad \text{with} \quad \begin{cases} v_\tau = 1 & \text{if } \xi_\tau > 0, \\ v_\tau \in [-1,1] & \text{if } \xi_\tau = 0, \\ v_\tau = -1 & \text{if } \xi_\tau < 0. \end{cases}$$

Then inequality (8.13c) can be represented as follows:

$$\sum_{\tau\in\mathcal{T}_h} \left( \int_\tau \bar{p}_h\, dx\, dt + |\tau|(\rho\bar{\xi}_\tau + \mu\bar{v}_\tau) \right)(v_\tau - \bar{\xi}_\tau) \;\geq\; 0 \quad \text{for all } a \leq v_\tau \leq b,$$

which is recasted in the equivalent form

$$\left( \int_\tau \bar{p}_h\, dx\, dt + |\tau|(\rho\bar{\xi}_\tau + \mu\bar{v}_\tau) \right)(v_\tau - \bar{\xi}_\tau) \;\geq\; 0, \quad a \leq v_\tau \leq b, \tag{8.14}$$

for all $\tau \in \mathcal{T}_h$. Therefore, from (8.14) we deduce the projection representation formula [38]

$$\bar{\xi}_\tau = \mathbf{Proj}_{[a,b]}\left( -\frac{1}{\rho}\left( \frac{1}{|\tau|} \int_\tau \bar{p}_h\, dx\, dt + \mu\bar{v}_\tau \right) \right) \tag{8.15}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

for $\bar{z}_h$ on each element $\tau \in \mathcal{T}_h$. From this, we derive the following results:

$$\bar{\xi}_\tau = 0 \;\Leftrightarrow\; \frac{1}{|\tau|}\left|\int_\tau \bar{p}_h \, dx\, dt\right| \le \mu,$$

$$\bar{v}_\tau = \mathbf{Proj}_{[-1,1]}\left(-\frac{1}{\mu|\tau|}\int_\tau \bar{p}_h \, dx\, dt\right).$$

These results are analogous to the discretization approach proposed in [1, 10] for the optimal control of semilinear elliptic equations and in [15] of the Schlögl and FitzHugh–Nagumo systems. In fact, by a closer look on the projection formula, we have the following form of the approximation $\bar{z}_h$ of an optimal control:

$$\bar{\xi}_\tau = \mathbb{P}(\bar{p}_\tau) = \begin{cases} b & \text{on } \mathcal{A}_{b,\mathcal{T}_h} := \{\tau \in \mathcal{T}_h : \bar{p}_\tau < -\rho b - \mu\}, \\ -\frac{1}{\rho}(\bar{p}_\tau + \mu) & \text{on } \mathcal{I}_{+,\mathcal{T}_h} := \{\tau \in \mathcal{T}_h : -\rho b - \mu \le \bar{p}_\tau < -\mu\}, \\ 0 & \text{on } \mathcal{A}_{0,\mathcal{T}_h} := \{\tau \in \mathcal{T}_h : |\bar{p}_\tau| \le \mu\}, \\ -\frac{1}{\rho}(\bar{p}_\tau - \mu) & \text{on } \mathcal{I}_{-,\mathcal{T}_h} := \{\tau \in \mathcal{T}_h : \mu < \bar{p}_\tau \le -\rho a + \mu\}, \\ a & \text{on } \mathcal{A}_{a,\mathcal{T}_h} := \{\tau \in \mathcal{T}_h : \bar{p}_\tau > -\rho a + \mu\}, \end{cases} \tag{8.16}$$

where

$$\bar{p}_\tau = \frac{1}{|\tau|}\int_\tau \bar{p}_h \, dx\, dt.$$

Inserting (8.15) into (8.13a), we obtain an equivalent form of the discretized optimality system, which consists of discretized state and adjoint state equations. Namely, find $\bar{u}_h \in X_{0,h}$ and $\bar{p}_h \in X_{T,h}$ such that $(\bar{u}_h, \bar{p}_h)$ solves the coupled system

$$\begin{aligned} &\int_Q \partial_t u_h \, v_h \, dx\, dt + \int_Q \nabla_x u_h \cdot \nabla_x v_h \, dx\, dt + \int_Q R(u_h)\, v_h \, dx\, dt \\ &\quad - \sum_{\tau \in \mathcal{T}_h} \int_\tau \mathbb{P}(\bar{p}_\tau)\, v_h \, dx\, dt = 0 \quad \text{for all } v_h \in X_{0,h}, \end{aligned} \tag{8.17a}$$

$$\begin{aligned} &-\int_Q u_h \, q_h \, dx\, dt - \int_Q \partial_t p_h \, q_h \, dx\, dt + \int_Q \nabla_x p_h \cdot \nabla_x q_h \, dx\, dt \\ &\quad + \int_Q R'(u_h) p_h \, q_h \, dx\, dt = -\int_Q u_Q \, q_h \, dx\, dt \quad \text{for all } q_h \in X_{T,h}. \end{aligned} \tag{8.17b}$$

The convergence of a solution of the discretized optimality system to a solution of the associated continuous optimality system and the error analysis of our finite element approximation are beyond the scope of this work and will be studied elsewhere.

To solve the above discretized coupled nonlinear optimality system, we apply the semismooth Newton method as discussed in [31], where a generalized derivative needs

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

to be computed at each Newton iteration. In fact, each iteration turns out to be one step of a primal-dual active set strategy [22], which is analogous to the continuous case we discussed in Section 8.3: Given $(u_h^k, p_h^k)$, find $(\delta u_h, \delta p_h)$ such that

$$\begin{aligned}
&\int_Q \partial_t \delta u_h \, v_h \, dx\, dt + \int_Q \nabla_x \delta u_h \cdot \nabla_x v_h \, dx\, dt + \int_Q R'(u_h^k)\, \delta u_h \, v_h \, dx\, dt \\
&\quad - \sum_{\tau \in \mathcal{T}_h} \int_\tau \mathbb{P}'(\bar{p}_\tau^k) \delta p_h \, v_h \, dx\, dt \\
&= - \int_Q \partial_t u_h^k \, v_h \, dx\, dt - \int_Q \nabla_x u_h^k \cdot \nabla_x v_h \, dx\, dt + \int_Q R(u_h^k)\, v_h \, dx\, dt \\
&\quad + \sum_{\tau \in \mathcal{T}_h} \int_\tau \mathbb{P}(\bar{p}_\tau^k)\, v_h \, dx\, dt
\end{aligned}$$

and

$$\begin{aligned}
&- \int_Q \delta u_h \, q_h \, dx\, dt - \int_Q \partial_t \delta p_h \, q_h \, dx\, dt + \int_Q \nabla_x \delta p_h \cdot \nabla_x q_h \, dx\, dt \\
&\quad + \int_Q R'(u_h^k) \delta p_h \, q_h \, dx\, dt + \int_Q R''(u_h^k) p_h^k \delta u_h \, q_h \, dx\, dt \\
&= \int_Q u_h^k \, q_h \, dx\, dt + \int_Q \partial_t p_h^k \, q_h \, dx\, dt - \int_Q \nabla_x p_h^k \cdot \nabla_x q_h \, dx\, dt \\
&\quad - \int_Q R'(u_h^k) p_h^k q_h \, dx\, dt - \int_Q u_Q \, q_h \, dx\, dt,
\end{aligned}$$

and $u_h^{k+1} = u_h^k + \omega \delta u_h$ and $p_h^{k+1} = p_h^k + \omega \delta p_h$ with some damping parameter $\omega \in (0,1]$, where $\bar{p}_\tau^k = \frac{1}{|\tau|} \int_\tau \bar{p}_h^k \, dx\, dt$.

## 8.5 Numerical experiments

For the two numerical examples considered in this section, we set $\Omega = (0,1)^2$, $T = 1$, and therefore $Q = (0,1)^3$. Using an octasection-based refinement [4], we uniformly decompose the space-time cylinder $Q$ into tetrahedral elements, i. e., the mesh size $h = 1/2, 1/4, \ldots$, until we reach $h = 1/128$. Therefore, the total numbers of degrees of freedom for the coupled state and adjoint state equations are 4,293,378. For the second example, we also use an adaptive refinement procedure driven by a residual-based error indicator for the coupled state and adjoint state system similar to that developed for the state equation in [34, 36]. More precisely, on each element $\tau \in \mathcal{T}_h$, we define the

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

local residuals

$$R_u^\tau(u_h,p_h) := [\bar{z}_h - \partial_t u_h + \Delta u_h - R(u_h)]|_\tau, \tag{8.18a}$$
$$R_p^\tau(u_h,p_h) := [-u_Q + u_h + \partial_t p_h + \delta p_h - R'(u_h)p_h]|_\tau, \tag{8.18b}$$

and on the common boundary $\gamma$ shared by the element $\tau$ and its neighboring element $\tau'$, we define the jumps of the normal fluxes

$$J_u^\gamma(u_h,p_h) := [n_x \cdot \nabla_x u_h + n_x' \cdot \nabla_x u_h']|_\gamma, \tag{8.19a}$$
$$J_p^\gamma(u_h,p_h) := [n_x \cdot \nabla_x p_h + n_x' \cdot \nabla_x p_h']|_\gamma, \tag{8.19b}$$

for both the state and adjoint state, respectively, where $n_x$ and $n_x'$ are the spatial components of the unit outward normal vectors to the inner boundaries $\gamma \subset \partial\tau$ and $\gamma \subset \partial\tau'$, respectively, and $u_h'$ and $p_h'$ are the discretized state and adjoint state solutions on the neighboring element $\tau'$. Then the local error indicators for the state and costate equations on each element $\tau$ are defined as

$$\eta_u^\tau(u_h,p_h) = \{c_1 h_\tau^2 \|R_u^\tau(u_h,p_h)\|_{L^2(\tau)}^2 + c_2 h_\tau \|J_u(u_h,p_h)\|_{L^2(\partial\tau)}^2\}^{\frac{1}{2}}, \tag{8.20a}$$
$$\eta_p^\tau(u_h,p_h) = \{c_3 h_\tau^2 \|R_p^\tau(u_h,p_h)\|_{L^2(\tau)}^2 + c_4 h_\tau \|J_p(u_h,p_h)\|_{L^2(\partial\tau)}^2\}^{\frac{1}{2}} \tag{8.20b}$$

with suitably chosen positive constants $c_i, i = 1,\dots,4$, which may depend on the model problem and domain shape (for simplicity, we choose $c_i = 1$ in our numerical experiments), and $h_\tau$ being the diameter of the element $\tau$. The total error indicator for our coupled nonlinear optimality system is then given by

$$\eta_{u,p}^\tau(u_h,p_h) = \eta_u^\tau(u_h,p_h) + \eta_p^\tau(u_h,p_h). \tag{8.21}$$

We perform our numerical tests on a desktop with Intel@ Xeon@ Processor E5-1650 v4 (15 MB Cache, 3.60 GHz, and 64 GB memory). For the nonlinear first-order necessary optimality system, we use the relative residual error $10^{-5}$ as a stopping criterion in the semismooth Newton iteration, whereas the algebraic multigrid preconditioned GMRES solver for the linearized system at each Newton iteration is stopped after a residual error reduction by $10^{-6}$; cf. also [36].

Presenting our numerical results, for simplicity, we will speak of the *sparse optimal control*. The term *computed numerical approximation of a sparse optimal control* would be more precise. We even assume that only one sparse optimal control exists, which is theoretically not sure. However, we observed only one solution of our discretized optimality systems.

### 8.5.1 Moving target (Example 1)

In the first numerical test, we use a benchmark example to illustrate that our space-time finite element method is able to recover an optimal control that is sparse in space

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

and time simultaneously. Therefore, we use the moving target

$$u_Q(x,t) = \exp(-20(x_1-0.2)^2+(x_2-0.2)^2+(t-0.2)^2) + \exp(-20(x_1-0.7)^2+(x_2-0.7)^2+(t-0.9)^2),$$

which is adapted from an example constructed in [11] to compare different sparsity properties; see an illustration of the target at times $t = 0.5$, 0.55, and 0.75 in Figure 8.1. The same desired state was also used in the numerical test for spatially directional sparse control in [13, 14]. The parameters in the optimal control problem are $\rho = 10^{-4}$, $\mu = 0.004$, $a = -10$, and $b = 20$. For the nonlinear reaction term in the state equation, we set $R(u) = u(u-0.25)(u+1)$. We use homogeneous initial and Dirichlet boundary conditions for the state equation.

To reach the relative residual error $10^{-5}$ for the nonlinear first order necessary optimality system, we needed 21 and 37 Newton iterations for the nonsparse and sparse optimal controls, respectively; see Figure 8.2. We clearly see the superlinear convergence of the semismooth Newton method after a few iterations when it reaches a sufficiently good approximation to the finite element solution of the nonlinear optimality

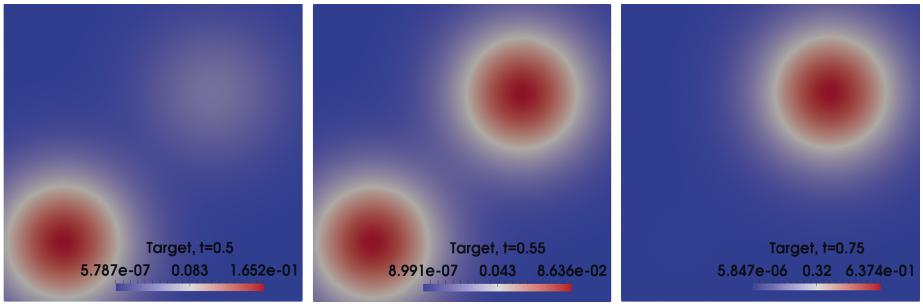



**Figure 8.1:** Example 1, plots of the target at times $t = 0.5,\ 0.55,\ 0.75$ for the moving target example.

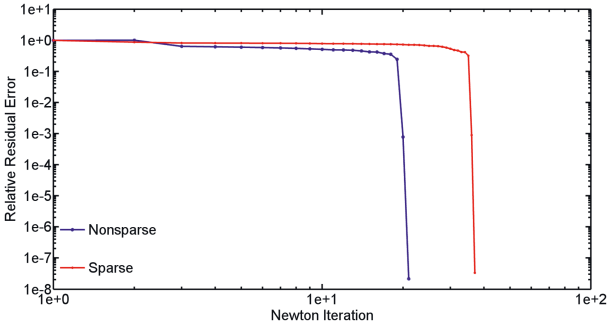



**Figure 8.2:** Example 1, relative residual error reduction in the semismooth Newton method for the nonsparse and sparse optimal controls in the moving target example.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

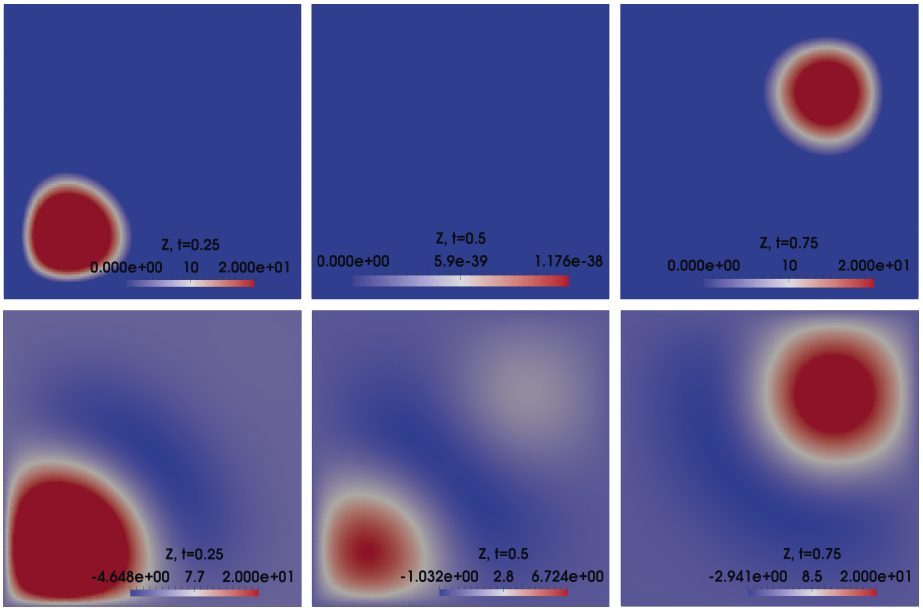



**Figure 8.3:** Example 1, Comparisons of sparse (up) and nonsparse optimal controls (down) at times $t = 0.25,\ 0.5,\ 0.75$ to illustrate spatial sparsity.

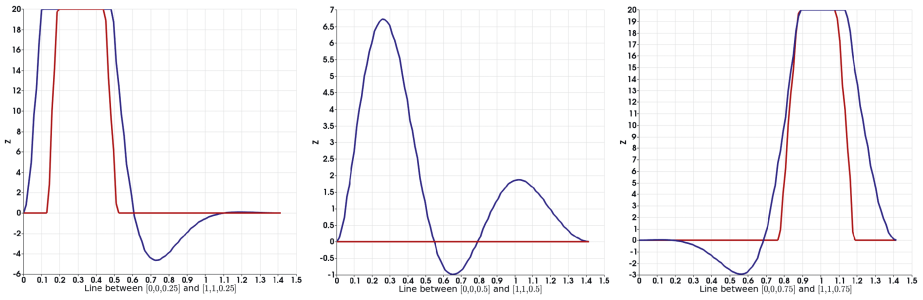



**Figure 8.4:** Example 1, comparisons of sparse (red) and nonsparse (blue) optimal controls along the line between $[0, 0, 0.25]$ and $[1, 1, 0.25]$, between $[0, 0, 0.5]$ and $[1, 1, 0.5]$, and between $[0, 0, 0.75]$ and $[1, 1, 0.75]$ for the moving target example.

system. The total system assembling and solving time is about 4 and 4.7 hours on the desktop computer, respectively.

Comparisons of sparse and nonsparse optimal controls at different times $t = 0.25$, 0.5, and 0.75 are displayed in Figure 8.3. We observe spatial sparsity at each time. The control even totally vanishes at the time $t = 0.5$. The difference between the sparse and nonsparse controls can be seen by a closer look at them along the diagonal of the spatial domain at different times $t = 0.25$, 0.5, and 0.75 as illustrated in Figure 8.4. Moreover, we visualize the sparse and nonsparse controls on the cutting plane $y = x$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

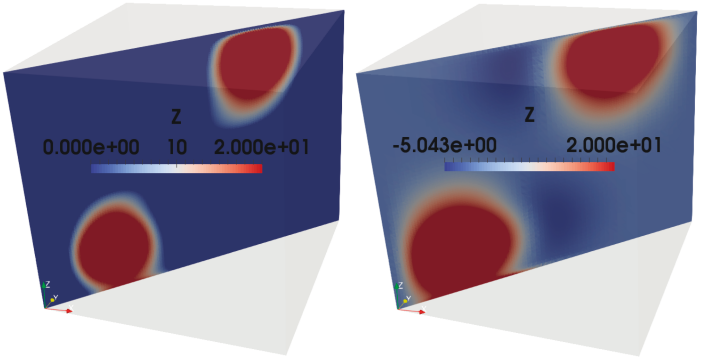



**Figure 8.5:** Example 1, Comparisons of sparse (left) and nonsparse optimal controls (right) on the plane $y = x$, $t \in (0, 1)$ to illustrate temporal sparsity.

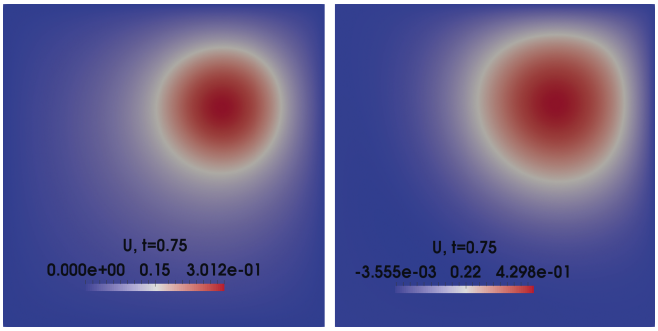



**Figure 8.6:** Example 1, Comparisons of states associated with sparse (left) and nonsparse optimal controls (right) at time $t = 0.75$ for the moving target example.

(with $z$ indicating the temporal direction) in the space-time domain; cf. Figure 8.5. It is also easy to see the temporal sparsity. The associated states with and without sparse control at time $t = 0.75$ are displayed in Figure 8.6.

In this example, we have seen that the $L^1$ cost functional promotes spatial and temporal sparsity in combination with our space-time method. These results are comparable to those of other authors, who used different discretization techniques. For example, in comparison with the sparse optimal control in the top right plot of Figure 1 in [11] (in one space dimension), we observe a similar sparsity of our optimal control.

## 8.5.2 Turning wave target (Example 2)

In the second test, we aim to recover optimal sparse controls for a parabolic equation with controlled turning wave fronts as studied in [15]. Therein, more involved

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

sparse control of the Schlögl and FitzHugh–Nagumo systems was considered, which was handled by a nonlinear conjugate gradient optimization method along with a conventional time stepping method. For a comparison, we consider the target

$$u_Q(x,t) = \left(1.0 + \exp\left(\frac{\cos(g(t))(\frac{70}{3} - 70x_1) + \sin(g(t))(\frac{70}{3} - 70x_2)}{\sqrt{2}}\right)\right)^{-1} + \left(1.0 + \exp\left(\frac{\cos(g(t))(70x_1 - \frac{140}{3}) + \sin(g(t))(70x_2 - \frac{140}{3})}{\sqrt{2}}\right)\right)^{-1} - 1,$$

where $g(t) = \frac{2\pi}{3}\min\{\frac{3}{4}, t\}$. This is an adapted version of the turning wave example considered in [15]. The wave front turns 90 degrees from time $t = 0$ to $t = 0.75$ and remains fixed after $t = 0.75$; see the target at $t = 0$, 0.25, 0.5, and 0.75 as illustrated in Figure 8.7.

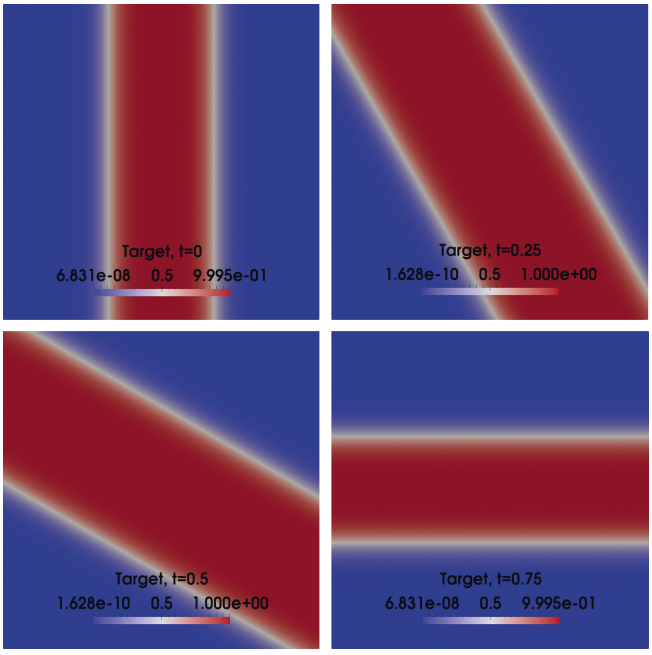



**Figure 8.7:** Example 2, plots of the target at time $t = 0,\ 0.25,\ 0.5,\ 0.75$ for the turning wave example.

The nonlinear reaction term is given by $R(u) = u(u-0.25)(u+1)$. We use the initial data

$$u_0(x) = \left(1 + \exp\left(\frac{\frac{70}{3} - 70x_1}{\sqrt{2}}\right)\right)^{-1} + \left(1 + \exp\left(\frac{70x_1 - \frac{140}{3}}{\sqrt{2}}\right)\right)^{-1} - 1$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

on $\Sigma_0$ and the homogeneous Neumann boundary condition on $\Sigma$ for the state. As parameters, we use $\rho = 10^{-6}$ and $\mu = 10^{-4}$ for the sparse case and $\rho = 10^{-6}$ and $\mu = 0$ for the nonsparse case. The bounds $a = -100$ and $b = 100$ are set for both cases.

To solve the nonlinear first-order necessary optimality system, we needed 7 and 35 semismooth Newton iterations for the nonsparse and sparse optimal controls, respectively; see Figure 8.8. The total system assembling and solving time is about 3 and 12.7 hours on the desktop computer, respectively. We again observe the superlinear convergence of the semismooth Newton method after it reaches a sufficiently good approximation to the finite element solution of the nonlinear optimality system.

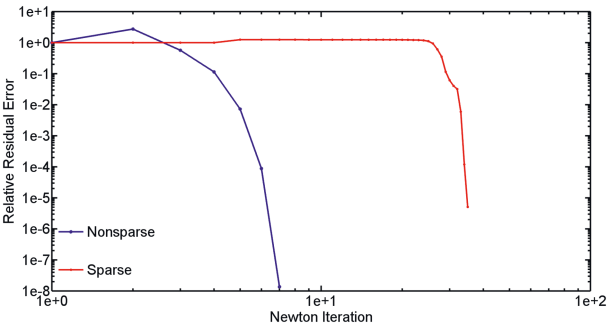



**Figure 8.8:** Example 2, relative residual error reduction in the semismooth Newton method for the nonsparse and sparse controls in the turning wave example.

The numerical solutions of sparse and nonsparse optimal controls and associated optimal states are illustrated in Figure 8.9. We clearly see a certain sparsity of our optimal sparse control as compared to pure $L^2$-regularization, without too much precision loss of the associated state to the target. A closer look on the computed sparse and nonsparse controls along the selected lines in the spatial domain at different times confirms that the computed discretized sparse control exhibits sparsity with respect to the spatial direction; see Figure 8.10. Similar sparse controls have been achieved in [15] using different approaches.

Instead of an uniform refinement, we may also adopt an adaptive strategy. In this way, local refinements are made in the region where the solution shows a more local character, whereas coarser meshes appear in the other region. For example, in the optimal sparse control case ($\rho = 10^{-6}$, $\mu = 10^{-4}$), we start from an initial mesh with 729 grid points, 9 points in each spatial and temporal direction. We use the residual-based error indicator $\eta^{\tau}_{u,p}(u_h, p_h)$, provided in (8.21), for the coupled state and adjoint state system to guide our adaptive mesh refinement. After the 6th adaptive octasection refinement [4], the mesh contains 1,053,443 grid points; see the adaptive space-time mesh and the meshes on the cutting planes at different times in Figure 8.11. As we observe, the adaptive refinements follow the rotation of the wave front of the state.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

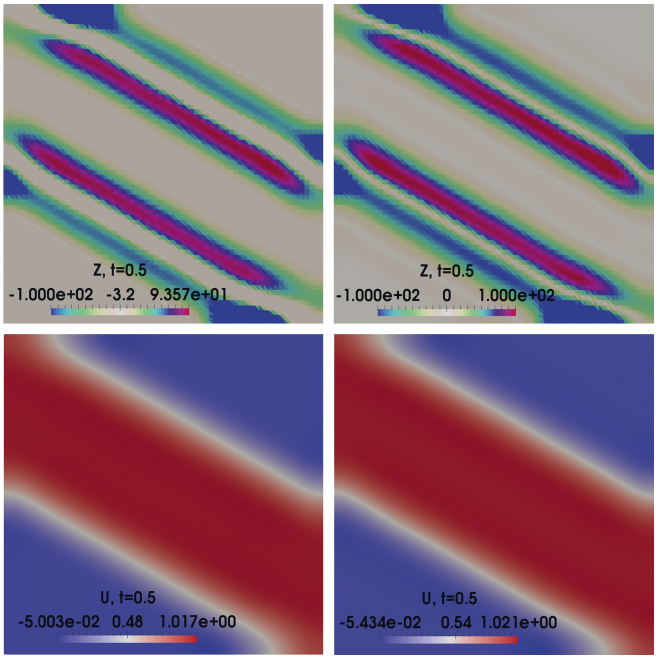



**Figure 8.9:** Example 2, plots of the sparse (left) and nonsparse (right) optimal controls (in the first row) and the associated states (in the second row) at time $t = 0.5$ for the turning wave example.

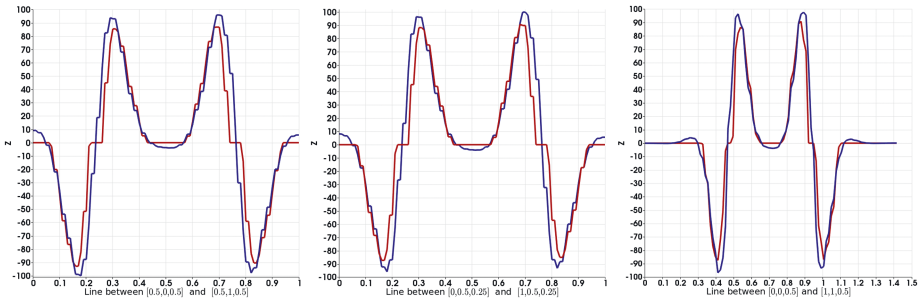



**Figure 8.10:** Example 2, comparisons of sparse (red) and nonsparse (blue) optimal controls along the lines between $[0.5, 0, 0.5]$ and $[0.5, 1, 0.5]$, between $[0, 0.5, 0.25]$ and $[1, 0.5, 0.25]$, and between $[0, 0, 0.5]$ and $[1, 1, 0.5]$ for the turning wave example.

For this example, we have observed that the $L^1$ cost functional promotes sparsity in combination with our space-time method, in particular, mainly in the spatial direction. This is comparable to the results in [15].

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

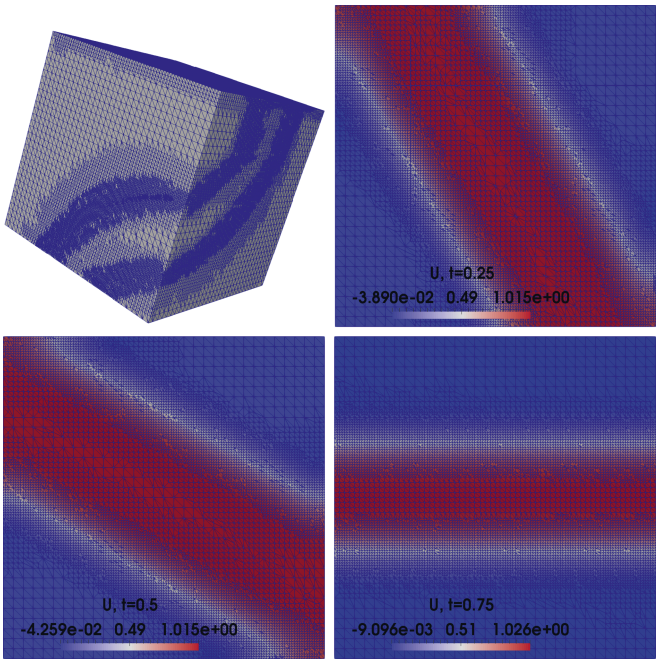



**Figure 8.11:** Example 2, plots of the adaptive space-time mesh (top-left) at the 6th step, and the meshes on the cutting planes for the times $t = 0.25$, 0.5, and 0.75.

## 8.6 Conclusions

In this work, we have considered the space-time Petrov–Galerkin finite element method on fully unstructured simplicial meshes for semilinear parabolic optimal sparse control problems. The objective functional involves the well-known $L^1$-norm of the control in addition to the standard $L^2$-regularization term. The proposed method is able to capture spatio-temporal sparsity, which has been confirmed by our numerical experiments. A rigorous convergence and error analysis of our space-time Petrov–Galerkin finite element methods for such an optimal sparse control problems is left for future work.

## Bibliography

[1] N. Arada, E. Casas, and F. Tröltzsch. Error estimates for the numerical approximation of a semilinear elliptic control problem. *Comput. Optim. Appl.*, 23:201–229, 2002.

[2] R. E. Bank, P. S. Vassilevski, and L. T. Zikatanov. Arbitrary dimension convection–diffusion schemes for space-time discretizations. *J. Comput. Appl. Math.*, 310:19–31, 2017.

[3] M. Behr. Simplex space-time meshes in finite element simulations. *Int. J. Numer. Methods Fluids*, 57:1421–1434, 2008.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[4] J. Bey. Tetrahedral grid refinement. *Computing*, 55:355–378, 1995.
[5] A. C. Boulanger and P. Trautmann. Sparse optimal control of the KdV–Burgers equation on a bounded domain. *SIAM J. Control Optim.*, 55(6):3673–3706, 2017.
[6] D. Braess. *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*. Cambridge University Press, 2007.
[7] E. Casas. A review on sparse solutions in optimal control of partial differential equations. *SeMA*, 74:319–344, 2017.
[8] E. Casas, C. Clason, and K. Kunisch. Parabolic control problems in measure spaces with sparse solutions. *SIAM J. Control Optim.*, 51(1):28–63, 2013.
[9] E. Casas, R. Herzog, and G. Wachsmuth. Approximation of sparse controls in semilinear equations by piecewise linear functions. *Numer. Math.*, 122:645–669, 2012.
[10] E. Casas, R. Herzog, and G. Wachsmuth. Optimality conditions and error analysis of semilinear elliptic control problems with $L^1$ cost functional. *SIAM J. Optim.*, 22(3):795–820, 2012.
[11] E. Casas, R. Herzog, and G. Wachsmuth. Analysis of spatio-temporally sparse optimal control problems of semilinear parabolic equations. *ESAIM Control Optim. Calc. Var.*, 23(1):263–295, 2017.
[12] E. Casas and K. Kunisch. Parabolic control problems in space-time measure spaces. *ESAIM Control Optim. Calc. Var.*, 22(2):355–370, 2016.
[13] E. Casas, M. Mateos, and A. Rösch. Finite element approximation of sparse parabolic control problems. *Math. Control Relat. Fields*, 7(3):393–417, 2017.
[14] E. Casas, M. Mateos, and A. Rösch. Improved approximation rates for a parabolic control problem with an objective promoting directional sparsity. *Comput. Optim. Appl.*, 70:239–266, 2018.
[15] E. Casas, C. Ryll, and F. Tröltzsch. Sparse optimal control of the Schlögl and FitzHugh–Nagumo systems. *Comput. Methods Appl. Math.*, 13(4):415–442, 2013.
[16] E. Casas, C. Ryll, and F. Tröltzsch. Second order and stability analysis for optimal sparse control of the FitzHugh–Nagumo equation. *SIAM J. Control Optim.*, 53(4):2168–2202, 2015.
[17] E. Casas and E. Zuazua. Spike controls for elliptic and parabolic PDEs. *Syst. Control Lett.*, 62(4):311–318, 2013.
[18] M. J. Gander. 50 years of time parallel integration. In T. Carraro, M. Geiger, S. Körkel, and R. Rannacher, editors, *Multiple Shooting and Time Domain Decomposition*, pages 69–114. Springer Verlag, Heidelberg, Berlin, Cham, 2015.
[19] M. J. Gander and M. Neumüller. Analysis of a new space-time parallel multigrid algorithm for parabolic problems. *SIAM J. Sci. Comput.*, 38(4):A2173–A2208, 2016.
[20] R. Herzog, J. Obermeier, and G. Wachsmuth. Annular and sectorial sparsity in optimal control of elliptic equations. *Comput. Optim. Appl.*, 62(1):157–180, 2015.
[21] R. Herzog, G. Stadler, and G. Wachsmuth. Directional sparsity in optimal control of partial differential equations. *SIAM J. Control Optim.*, 50(2):943–963, 2012.
[22] K. Ito and K. Kunisch. *Lagrange multiplier approach to variational problems and applications*, volume 15 of Advances in Design and Control. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.
[23] V. Karyofylli, L. Wendling, M. Make, N. Hosters, and M. Behr. Simplex space-time meshes in thermally coupled two-phase flow simulations of mold filling. *Comput. Fluids*, 192:104261, 2019.
[24] K. Kunisch, K. Pieper, and B. Vexler. Measure valued directional sparsity for parabolic optimal control problems. *SIAM J. Control Optim.*, 52(5):3078–3108, 2014.
[25] U. Langer, Neumüller M., and Schafelner A. Space-time finite element methods for parabolic evolution problems with variable coefficients. In T. Apel, U. Langer, A. Meyer, and O. Steinbach, editors, *Advanced Finite Element Methods with Applications: Selected Papers from the 30th Chemnitz Finite Element Symposium 2017*, pages 247–275. Springer International Publishing, Cham, 2019.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[26] U. Langer, O. Steinbach, F. Tröltzsch, and H. Yang. Unstructured space-time finite element methods for optimal control of parabolic equations. *SIAM J. Sci. Comput.*, 43(2):A744–A771, 2021.

[27] U. Langer and H. Yang. BDDC preconditioners for a space-time finite element discretization of parabolic problems. In R. Haynes, S. MacLachlan, X. Cai, L. Halpern, H. Kim, A. Klawonn, and O. Widlund, editors, *Domain Decomposition Methods in Science and Engineering XXV*, volume 138 of Lecture Notes in Computational Science and Engineering, pages 367–374. Springer International Publishing, Cham, 2020.

[28] C. Li and G. Stadler. Sparse solutions in optimal control of PDEs with uncertain parameters: the linear case. *SIAM J. Control Optim.*, 57(1):633–658, 2019.

[29] A. Nägel, D. Logashenko, J. B. Schroder, and U. M. Yang. Aspects of solvers for large-scale coupled problems in porous media. *Transp. Porous Media*, 130:363–390, 2019.

[30] C. Ryll, J. Löber, S. Martens, H. Engel, and F. Tröltzsch. Analytical, optimal, and sparse optimal control of traveling wave solutions to reaction–diffusion systems. In E. Schöll, S. H. L. Klapp, and P. Hövel, editors, *Control of Self-Organizing Nonlinear Systems*, pages 189–210. Springer International Publishing, Cham, 2016.

[31] G. Stadler. Elliptic optimal control problems with $L^1$-control cost and applications for the placement of control devices. *Comput. Optim. Appl.*, 44:159–181, 2009.

[32] O. Steinbach. *Numerical approximation methods for elliptic boundary value problems*. Springer-Verlag, New York, 2008.

[33] O. Steinbach. Space-time finite element methods for parabolic problems. *Comput. Methods Appl. Math.*, 15:551–566, 2015.

[34] O. Steinbach and H. Yang. Comparison of algebraic multigrid methods for an adaptive space-time finite-element discretization of the heat equation in 3D and 4D. *Numer. Linear Algebra Appl.*, 25(3):e2143, 2018.

[35] O. Steinbach and H. Yang. A space-time finite element method for the linear bidomain equations. In T. Apel, U. Langer, A. Meyer, and O. Steinbach, editors, *Advanced Finite Element Methods with Applications: Selected Papers from the 30th Chemnitz Finite Element Symposium 2017*, pages 323–339. Springer International Publishing, Cham, 2019.

[36] O. Steinbach and H. Yang. Space-time finite element methods for parabolic evolution equations: discretization, a posteriori error estimation, adaptivity and solution. In O. Steinbach and U. Langer, editors, *Space-Time Methods: Application to Partial Differential Equations*. Radon Series on Computational and Applied Mathematics, pages 207–248, de Gruyter, Berlin, 2019.

[37] I. Toulopoulos. Space-time finite element methods stabilized using bubble function spaces. *Appl. Anal.*, 99(7):1153–1170, 2020.

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

[39] M. Ulbrich. *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, volume 11 of MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, 2011.

[40] M. Uzunca, T. Küçükseyhan, H. Yücel, and B. Karasözen. Optimal control of convective FitzHugh–Nagumo equation. *Comput. Math. Appl.*, 73(9):2151–2169, 2017.

[41] M. von Danwitz, V. Karyofylli, N. Hosters, and M. Behr. Simplex space-time meshes in compressible flow simulations. *Int. J. Numer. Methods Fluids*, 91(1):29–48, 2019.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Carolin Dirks and Benedikt Wirth

# 9 An adaptive finite element approach for lifted branched transport problems

**Abstract:** We consider the so-called branched transport and variants thereof in two space dimensions. In these models, we seek an optimal transportation network for a given mass transportation task. In two space dimensions, they are closely connected to Mumford–Shah-type image processing problems, which in turn can be related to certain higher-dimensional convex optimization problems via so-called functional lifting. We examine the relation between these different models and exploit it to solve the branched transport model numerically via convex optimization. To this end, we develop an efficient numerical treatment based on a specifically designed class of adaptive finite elements. This method allows the computation of finely resolved optimal transportation networks despite the high dimensionality of the convex optimization problem and its complicated set of nonlocal constraints. In particular, by design of the discretization the infinite set of constraints reduces to a finite number of inequalities.

**Keywords:** branched transport, urban planning, functional lifting, calibration, finite elements, Mumford–Shah problem

**MSC 2010:** 49Q10, 49M20, 49M29

## 9.1 Introduction

The concept of optimal transport essentially is the following. Given two nonnegative (probability) measures, a material source $\mu_+$, and a material sink $\mu_-$, say in $\mathbf{R}^d$, we need to transport the material from $\mu_+$ to $\mu_-$ at minimal cost. In *classical* optimal transport, this amounts to deciding how much mass goes from whence to where. The original formulation by Monge from the 18th century does so by introducing a transport map $T : \mathbf{R}^d \to \mathbf{R}^d$, encoding that material in $x$ should be moved to $T(x)$, and identifying its

**Acknowledgement:** The work was supported by the Alfried Krupp Prize for Young University Teachers awarded by the Alfried Krupp von Bohlen und Halbach-Stiftung and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy through the Cluster of Excellence "Mathematics Münster: Dynamics – Geometry – Structure" (EXC 2044 – 390685587) at the University of Münster and through the DFG-grant WI 4654/1-1 within the Priority Program 1962.

**Carolin Dirks, Benedikt Wirth,** University of Münster, Applied Mathematics Münster, Einsteinstr. 62, D-48149 Münster, Germany, e-mails: mail@carolin-dirks.de, Benedikt.Wirth@uni-muenster.de

https://doi.org/10.1515/9783110695984-009

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

cost as

$$\int_{\mathbf{R}^d} c(x, T(x))\, \mathrm{d}\mu_+(x) \tag{9.1}$$

with $c(x,y)$ the cost for transporting one unit mass from $x$ to $y$ (typical choices are $c(x,y) = |x-y|^p$ for $p \geq 1$, leading to the so-called Wasserstein-$p$ transport). The task now is to find that transport map $T$ with minimal cost among all transport maps that push $\mu_+$ forward onto $\mu_-$, in other words, which move the material from $\mu_+$ in such a fashion that after the transport, it is distributed according to $\mu_-$. Monge's formulation turns out to be not well-posed in general since it does not allow for material in $x$ being transported to two or more positions. Kantorovich remedied this in the last century by replacing transport maps with so-called transport plans, probability measures $\pi$ on $\mathbf{R}^d \times \mathbf{R}^d$ with, roughly speaking, $\pi(x,y)$ having the interpretation of how much mass is transported from $x$ to $y$. In terms of a transport plan $\pi$, the cost turns into

$$\int_{\mathbf{R}^d \times \mathbf{R}^d} c(x,y)\, \mathrm{d}\pi(x,y), \tag{9.2}$$

to be minimized among all nonnegative measures satisfying $\pi(\cdot, \mathbf{R}^d) = \mu_+$ and $\pi(\mathbf{R}^d, \cdot) = \mu_-$. This formulation is well-posed and even represents a convex linear program in $\pi$, which via convex duality techniques allows for a large variety of equivalent problem reformulations. In the particular case of Wasserstein-1 transport, in which the cost $c(x,y)$ is proportional to the transport distance, one such reformulation (the so-called Beckmann formulation) considers the material flux $\mathcal{F}$ as variable. The flux $\mathcal{F}$ is a vector-valued Radon measure describing the material flow from the sources to the sinks, and its cost turns out to be simply its total mass or total variation,

$$|\mathcal{F}|(\mathbf{R}^d), \tag{9.3}$$

to be minimized among all vector-valued Radon measures satisfying $\operatorname{div} \mathcal{F} = \mu_+ - \mu_-$ (simply think of $\mathcal{F}$ like a smooth vector field, whose divergence is known to equal sources minus sinks). As a preparation for the introduction of the so-called *branched transport* below, note that mass fluxes describing the transport of material from $\mu_+$ to $\mu_-$ can be decomposed into fluxes along one-dimensional paths (just think of the trajectories of the transported particles). If many particles follow the same trajectory, then $\mathcal{F}$ concentrates on this one-dimensional trajectory, and thus it is not unnatural to decompose $\mathcal{F}$ into $\mathcal{F} = \theta(\mathcal{H}^1 \llcorner S) + \mathcal{F}^d$ for $\mathcal{H}^1 \llcorner S$ the restriction of the one-dimensional Hausdorff measure to a one-dimensional (or, more precisely, countable $\mathcal{H}^1$-rectifiable) set $S \subset \mathbf{R}^d$, $\theta : S \to \mathbf{R}^d$ a flow strength tangent to $S$ (the amount of particles times the unit tangent vector), and $\mathcal{F}^d$ the nonconcentrated or diffuse remainder of $\mathcal{F}$. In this

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

case the cost can be written as (see [9] for a rigorous derivation)

$$|\mathcal{F}|(\mathbf{R}^d) = \int_S |\theta(x)| \, \mathrm{d}\mathcal{H}^1(x) + |\mathcal{F}^d|(\mathbf{R}^d). \tag{9.4}$$

In all of the above, the cost per transport distance is proportional to the transported mass. This holds in particular also for Wasserstein-1 transport (recall that $|\theta|$ represents the amount of particles traveling through $x$). During the past two decades, a class of variants of Wasserstein-1 transport have been developed, whose underlying cost functionals have the feature that the cost $\tau(m)$ per transport distance is no longer proportional to the amount of transported mass $m$, but rather only subadditive in $m$. In terms of the above formulations, the cost of these new models is given by

$$\int_S \tau(|\theta(x)|) \, \mathrm{d}\mathcal{H}^1(x) + \tau'(0)|\mathcal{F}^d|(\mathbf{R}^d), \tag{9.5}$$

to be minimized for $S$, $\theta$, and $\mathcal{F}^d$ under the constraint $\operatorname{div}(\theta(\mathcal{H}^1 \llcorner S) + \mathcal{F}^d) = \mu_+ - \mu_-$ (just like before in the Wasserstein-1 case). This subadditive cost function penalizes transport of small masses disproportionately stronger and thus promotes mass aggregation as can be illustrated by the following example. Rather than transporting mass $m_1$ and mass $m_2$ along two neighboring paths of length $l$, which by the above would cost $l\tau(m_1) + l\tau(m_2)$, it is cheaper to first coalesce both paths so that both masses are transported together, yielding $l\tau(m_1 + m_2)$ (which in case of strict subadditivity of $\tau$ is smaller). As a consequence, network-like structures emerge during the optimization, along which as many particles travel together as possible. The resulting networks exhibit a complicated branching structure, where the grade of ramification and the network geometry are controlled by the precise form of the cost functional. Particular instances of this model class include the so-called branched transport [23, 38], urban planning [6], and the Steiner tree problem [20] (note that there is a large variety of possible model formulations, which in the end turn out to be equivalent; see [9] and the references therein). There exist a variety of interesting applications such as the optimization of communication or public transportation networks [11, 19] or the understanding of vascular structures in plants and animals [10, 39, 42], to name just a few.

Typically, the corresponding energy landscape is highly nonconvex. Consequently, the identification and construction of a globally optimal transportation network is a challenging task. In this work, we exploit a connection of the two-dimensional transportation network problem to convex image processing methods to compute globally optimal network geometries numerically. We already made use of this connection in previous work [7] to prove lower bounds on the transportation cost and to perform preliminary numerical simulations; however, since our sole interest were lower bounds, we had neither fully understood the underlying connection nor

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

come up with an efficient, tailored numerical scheme. In contrast, in the present work, our focus is on numerically solving two-dimensional branched transport problems. For this, we will prove the equivalence of the original transportation network problem to a sequence of models leading to a convex image inpainting problem (the only gap in this sequence of equivalences will be a classical relaxation step for the Mumford–Shah functional, whose tightness is unto this date not known to the best of our knowledge). Even though the final problem is convex, it features a high dimensionality and a huge number of constraints, which render its solution with standard methods infeasible. We thus proceed to design a particular adaptive discretization, which tremendously decreases the computational effort and thereby allows computation of highly resolved optimal transportation schemes.

### 9.1.1 Existing numerical methods for branched transport-type problems

To simulate optimal transportation networks, several approaches have been investigated in the literature. Based on the Eulerian formulation via mass fluxes, Xia [38, 41] introduced an initial approach for numerically finding an optimal graph between two measures. This local optimization technique was extended to a minimization algorithm in [40], which in several numerical examples with a single source point and a fixed number of $N$ sinks seems to yield almost optimal networks. It was shown in [40] and [41] that, although not necessarily leading to a global minimizer, this optimization algorithm provides an approximately optimal transport network and is applicable even in case of a large number of sinks ($N \approx 400$). Two heuristic approaches based on stochastic optimization techniques on graphs were presented in [25] and [30]. As before, these methods are capable of providing almost optimal network structures but cannot guarantee global optimality either. The limit case of the Steiner tree problem, where the transport cost is independent of the amount of transported mass, was treated more extensively in the literature. Due to the independence of the transported mass, there exist very efficient algorithms in a planar geometry providing a globally optimal Steiner tree (see, for instance, the GeoSteiner method [21] or Melzak's full Steiner tree algorithm [26]). For more than two space dimensions, there exist fewer approaches, which are less efficient; an overview of some methods for the Steiner tree problem in $n$ dimensions is provided in [15], where the main ideas trace back to [18, 20, 22, 35]. For the general transportation network problem, a widely used approach was inspired by elliptic approximations of free-discontinuity problems in the sense of Modica–Mortola and Ambrosio–Tortorelli via phase fields. In [12, 16, 17, 27, 29, 37], corresponding phase field approximations have been presented for the classical branched transport problem, the Steiner tree problem, a variant of the urban planning problem (which is piecewise linear in the amount of transported mass) or more general cost functions, however, all restricted to two space dimensions.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

### 9.1.2 Contributions of our work

In this work, we build on the approach introduced by [7], which consists in a novel reformulation of the optimal transportation network problem as a Mumford–Shah-type image inpainting problem in two dimensions. Roughly speaking, the optimal network is represented by the rotated gradient of a grey-value image of bounded variation. The resulting equivalent energy functional resembles the structure of the well-known Mumford–Shah functional [28], which in turn admits a convex higher-dimensional relaxation by a so-called functional lifting approach [1, 31].

In a little more detail, fix some domain $\Omega \subset \mathbf{R}^2$ and consider a source and sink $\mu_+$ and $\mu_-$ supported on the boundary $\partial\Omega$. We denote by $\tau(m)$ the cost for transporting mass $m$ along one unit distance. Describing the transportation network as a vector measure $\mathcal{F} \in \mathcal{M}(\overline{\Omega}; \mathbf{R}^2)$, the associated generalized branched transport cost can be defined as some functional $\mathcal{E}(\mathcal{F})$ depending on the choice of $\tau$. Now consider any vector measure $\mathcal{F} \in \mathcal{A}_{\mathcal{F}} = \{\mathcal{F} \in \mathcal{M}(\overline{\Omega}; \mathbf{R}^2) \mid \operatorname{div} \mathcal{F} = \mu_+ - \mu_-\}$. Rotating its vector component pointwise by $\frac{\pi}{2}$, it can be identified with the (distributional) gradient $Du$ of a function $u : \Omega \to \mathbf{R}$, which in mathematical image processing is typically interpreted as a grey-value image. This image $u$ will lie in a subset $\mathcal{A}_u \subset \mathrm{BV}(\Omega)$ of the space of functions of bounded variation, where $\mathcal{A}_u$ roughly denotes the set of all those images that correspond to a transportation network between $\mu_+$ and $\mu_-$. With such an identification of transportation networks and images, the generalized branched transport cost can be reformulated as the cost

$$\tilde{\mathcal{E}}(u) = \int_{S_u \cap \overline{\Omega}} \tau([u]) \, \mathrm{d}\mathcal{H}^1 + \tau'(0)|Du|(\Omega \setminus S_u) \tag{9.6}$$

of the associated image $u$ (where $S_u$ denotes the discontinuity set, and $[u]$ the jump of $u$, whereas $|Du|$ is the total-variation measure of $Du$). Writing $1_u$ for the characteristic function of the subgraph of $u$ and $D1_u$ for its gradient, which is a measure concentrated on the graph of $u$, Alberti et al. [1] suggested to rewrite $\tilde{\mathcal{E}}(u)$ as

$$\mathcal{G}(1_u) = \sup_{\phi \in \mathcal{K}} \int_{\overline{\Omega} \times \mathbf{R}} \phi \cdot \mathrm{d}D1_u \tag{9.7}$$

for some particular set $\mathcal{K}$ of three-dimensional vector fields depending on $\tau$. By convexifying the set of characteristic functions $1_u$ to a set $\mathcal{C}$ of more general functions $v : \Omega \times \mathbf{R} \to [0, 1]$ we finally arrive at a convex optimization problem, whose dual can be used to provide a lower bound. In summary, as proved rigorously in [7], we have

$$\begin{aligned} \inf_{\mathcal{F} \in \mathcal{A}_{\mathcal{F}}} \mathcal{E}(\mathcal{F}) &\geq \inf_{u \in \mathcal{A}_u} \tilde{\mathcal{E}}(u) \geq \inf_{u \in \mathcal{A}_u} \mathcal{G}(1_u) \geq \inf_{v \in \mathcal{C}} \mathcal{G}(v) \\ &\geq \sup_{\phi \in \mathcal{K}} \int_{\partial\Omega \times \mathbf{R}} 1_{u(\mu_+, \mu_-)} \phi \cdot n \, \mathrm{d}\mathcal{H}^2 - \int_{\Omega \times \mathbf{R}} \max\{0, \operatorname{div} \phi\} \, \mathrm{d}x \, \mathrm{d}s, \end{aligned} \tag{9.8}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

where $1_{u(\mu_+,\mu_-)}$ denotes a particular binary function defined on $\partial\Omega \times \mathbf{R}$. The left-hand side of the above is the original generalized branched transport problem. In [7], we used the right-hand side to prove lower bounds for $\mathcal{E}(\mathcal{F})$, and we furthermore discretized this three-dimensional convex optimization problem via a simple finite difference scheme and presented several simulation results for different scenarios.

From the viewpoint of numerics for branched transport problems, the results of [7] are unsatisfactory for two reasons: (i) The final convex optimization problem was only shown to be a lower bound, whose solutions might actually differ from the minima of the original problem; (ii) The employed numerical methods suffered from excessive memory and computation time requirements, rendering complex network optimizations infeasible. The contribution of the present work is to remedy these shortcomings:

- We prove equality for the whole above sequence of inequalities except for the third, $\inf_{u\in\mathcal{A}_u} \mathcal{G}(1_u) \geq \inf_{v\in\mathcal{C}} \mathcal{G}(v)$, which we only conjecture to be an equality (this is a particular instance of a long-standing, yet unsolved problem, for which we can only provide some discussion and numerical evidence). Note that although the equality $\inf_{u\in\mathcal{A}_u} \tilde{\mathcal{E}}(u) = \inf_{u\in\mathcal{A}_u} \mathcal{G}(1_u)$ might be considered known in the calculus of variations community (it is, for instance, stated as Remark 3.3 in the arXiv version of [1]), a rigorous proof was not available in the literature.
- For a set of simple example cases, we provide fluxes $\mathcal{F} \in \mathcal{A}_{\mathcal{F}}$ and vector fields $\phi \in \mathcal{K}$ for which the left- and right-hand sides in the above inequality coincide (such vector fields are known as calibrations). This serves the same two purposes as the original use of calibrations in [1] for the classical Mumford–Shah functional: It shows $\inf_{u\in\mathcal{A}_u} \mathcal{G}(1_u) = \inf_{v\in\mathcal{C}} \mathcal{G}(v)$ in various relevant cases, and it provides explicit optimality results for particular settings of interest.
- We develop a nonstandard finite element scheme, allowing an efficient treatment of the lifted branched transportation network problem and providing locally high resolution results. The main difficulties here lie in the high dimensionality due to the lifted dimension and in the suitable handling of the infinite number of nonlocal inequality constraints defining the set $\mathcal{K}$. The former difficulty is approached by the use of grid adaptivity, and the latter by a particular design of the finite element discretization.

The above-mentioned equalities are presented and proved in Theorems 9.2.1–9.2.3 within Section 9.2, which also contains the calibration examples. The tailored discretization and corresponding numerical algorithm are presented in Section 9.3 together with numerical results.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

### 9.1.3 Preliminaries

Let us briefly fix some notation. We denote by $\mathcal{L}^n$ the $n$-dimensional Lebesgue measure, by $\mathcal{H}^k$ the $k$-dimensional Hausdorff measure, and by $\delta_x$ the Dirac measure at a point $x \in \mathbf{R}^n$. The space of $\mathbf{R}^N$-valued Radon measures on $\overline{\Omega}$ for an open bounded domain $\Omega \subset \mathbf{R}^n$ is denoted by $\mathcal{M}(\overline{\Omega}; \mathbf{R}^N)$. For $N = 1$, we write $\mathcal{M}(\overline{\Omega})$ and define $\mathcal{M}_+(\overline{\Omega})$ as the set of nonnegative finite Radon measures on $\overline{\Omega}$. For a measure $\mathcal{F} \in \mathcal{M}(\overline{\Omega}; \mathbf{R}^N)$, the corresponding total-variation measure and the total-variation norm are denoted by $|\mathcal{F}|$ and $\|\mathcal{F}\|_{\mathcal{M}} = |\mathcal{F}|(\overline{\Omega})$, respectively. The Radon measures can be viewed as the dual to the space of continuous functions, and thus there is the corresponding notion of weak-* convergence, indicated by $\overset{*}{\rightharpoonup}$. For a measure space $(X, \mathcal{A}, \mu)$ and some $Y \subset X$ with $Y \in \mathcal{A}$, the restriction of the measure $\mu$ onto $Y$ is written as $\mu \llcorner Y(A) = \mu(A \cap Y)$ for all $A \in \mathcal{A}$. The Banach space of functions of bounded variation on $\Omega$, that is, functions $u$ in the Lebesgue space $L^1(\Omega)$ whose distributional derivative is a vector-valued Radon measure, is denoted $\mathrm{BV}(\Omega)$ with norm $\|u\|_{\mathrm{BV}} = \|u\|_{L^1} + \|Du\|_{\mathcal{M}}$. The Banach space of continuous $\mathbf{R}^N$-valued functions on $\overline{\Omega}$ is denoted by $C^0(\overline{\Omega}; \mathbf{R}^N)$, and the space of compactly supported smooth $\mathbf{R}^N$-valued functions on $\Omega$ by $C_0^\infty(\Omega; \mathbf{R}^N)$.

For a convex subset $C$ of a vector space $X$, we write the orthogonal projection of $x \in X$ onto $C$ as $\pi_C(x) = \arg\min_{y \in C} |x - y|$. The convex analysis indicator function of $C$ is defined by $\iota_C(x) = 0$ if $x \in C$ and $\iota_C(x) = \infty$ else.

In Table 9.1 we provide an alphabetical reference table of the symbols used throughout the paper.

Finally, let us comment on a few notions that have different meanings in different communities and thus should be clarified here. When we speak of a (transportation) *network*, we think of something like a pipe or street network that serves the purpose of transporting material from sources to sinks. Information on the network also includes the capacity of all its branches (its pipes or streets), that is, how much mass is carried by each of them. Mathematically, such a network will be expressed as a *mass flux* (Definition 9.2.4). The transportation networks or mass fluxes will typically exhibit one-dimensional branches, but in principle the material transport may (and sometimes does) happen in a more diffuse way (think of regions where cars may go off-road and take any path they want). The simplest networks can also be represented as an *embedded discrete graph* in the sense of discrete mathematics, that is, a collection of straight edges representing the network branches. This notion of a graph is used in Section 9.2.1, whereas in Section 9.2.2, we require the concept of the *graph of a function* to introduce a convexification of the branched transport problem. With the notion of a *measure*, we always refer to a Radon measure as introduced at the beginning of this section. Whereas material sources and sinks will be represented by *nonnegative measures*, mass fluxes will be *vector-valued measures*. Last, the notion of an *image* will always refer to its meaning in mathematical image processing, where it is nothing else but a real-valued function defined on some two-dimensional domain, which

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Table 9.1:** Summary of used symbols.

| | |
|---|---|
| $1_u$ | characteristic function of the subgraph of an image $u$ (see (9.20)) |
| $\mathcal{A}_{\mathcal{F}}$, $\mathcal{A}_u$ | sets of admissible fluxes and images (Definition 9.2.6) |
| $\mathcal{C}$ | domain of convex generalized branched transport cost (Definition 9.2.9) |
| $\chi_{\{u>s\}}$ | characteristic function of the $s$-superlevel set of a function $u$ (see (9.35)) |
| $\mathcal{D}$ | dual functional to convex generalized branched transport cost (Theorem 9.2.3) |
| $\mathcal{E}(\mathcal{F})$ | generalized branched transport cost functional (Definitions 9.2.3 and 9.2.5) |
| $\tilde{\mathcal{E}}(u)$ | image-based cost functional (Definition 9.2.7) |
| $\eta_T$ | elementwise adaptive refinement indicator (see (9.97)) |
| $\mathcal{F}$, $\mathcal{F}_G$, $\mathcal{F}_u$ | mass flux and mass flux associated with a graph $G$ or an image $u$ (Definitions 9.2.4 and 9.2.2 and (9.13)) |
| $\phi = (\phi^x, \phi^s)$ | dual variable of lifting formulations (Definitions 9.2.8 and 9.2.9) |
| $G$, $V(G)$, $E(G)$, $w$ | weighted directed graph and associated vertex and edge set and weight function (Definition 9.2.2) |
| $\mathcal{G}$, $\tilde{\mathcal{G}}$ | surface-based and convex generalized branched transport cost functionals (Definitions 9.2.8 and 9.2.9) |
| $J$, $g$, $h$ | Mumford–Shah functional with integrands $g$ and $h$ (see (9.19)) |
| $\mathcal{K}$, $\tilde{\mathcal{K}}$, $\hat{\mathcal{K}}$, $\mathcal{K}_x$, $\mathcal{K}_s$ | constraint sets of functional lifting formulations (Definition 9.2.8, Theorem 9.2.2, Remark 9.2.2, (9.95), (9.96)) |
| $M$ | total transported mass (Theorem 9.2.2) |
| $\mu_+$, $\mu_-$ | nonnegative Radon measures representing source and sink |
| $\mathcal{N}$, $\mathcal{N}'$, $\mathcal{N}''$ | nodes, nonhanging nodes, and nonhanging nontop nodes of grid (Definition 9.3.4 and (9.85)) |
| $\Omega$ | computational domain |
| $s$ | coordinate along additional lifting dimension |
| $S_u$, $\nu_u$, $[u]$ | the approximate discontinuity set of an image $u$, its unit normal, and discontinuity size (see (9.13)) |
| $S^1(\mathcal{T})$, $S^{0,1}(\mathcal{T})$ | finite element spaces of the discretization (see (9.83), (9.84)) |
| $\mathcal{T}$, $T$ | triangular prism grid and prism element (Definition 9.3.4) |
| $\tau$ | transportation cost (Definition 9.2.1) |
| $\tau^{\mathrm{bt}}$, $\tau^{\mathrm{up}}$, $\tau^{\mathrm{st}}$ | branched transport, urban planning, and Steiner tree cost (Example 9.2.1) |
| $u$, $u_{\mathcal{F}}$ | an image and the image associated with a mass flux (Section 9.2.2) |
| $u(\mu_+, \mu_-)$ | boundary values for images associated with transports from source $\mu_+$ to sink $\mu_-$ (Definition 9.2.6) |
| $V = B_1(\Omega)$ | 1-neighborhood of computational domain (Section 9.2.2) |
| $v$ | optimization variable of convex generalized branched transport cost (Definition 9.2.9) |
| $x$ | two-dimensional spatial coordinate |

is usually used as a model for grey-value images. Typically (but not necessarily), images are assumed to satisfy additional regularity such as being of bounded variation, and this will also be the case here. We will speak of an image rather than merely of a function since the techniques we will apply to these functions stem from the field of mathematical image processing and there are applied to images.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

## 9.2 Functional lifting of the generalized branched transport cost

Below we briefly recapitulate the Eulerian formulation of the generalized branched transport problem in Section 9.2.1, after which we introduce the reformulation as Mumford–Shah image inpainting problem and its convexification via functional lifting in Section 9.2.2. We will prove the equivalence of the different resulting formulations except for one relaxation step, whose implications can only be discussed. In particular, whereas the previous work only established the inequalities in (9.8), we here prove the opposite inequalities, thereby establishing for the first time a convex reformulation of (generalized) branched transport. We then use the convex optimization problem to show optimality of a few particular network configurations in Section 9.2.3, among them, an example of diffuse transport, whose global optimality cannot be shown with existing convexification or calibration techniques from the literature.

### 9.2.1 Generalized branched transport

In generalized branched transport models, the cost for transporting a lump of mass $m$ along one unit distance is described by a transportation cost $\tau(m)$. This transportation cost is taken to be subadditive, which encodes that transporting several lumps of mass together is cheaper than transporting each separately. (Two further natural requirements from an application viewpoint are monotonicity and lower semicontinuity.) For the purpose of this paper, we restrict ourselves to the class of concave transportation costs (note that any concave function $\tau$ with $\tau(0) = 0$ is subadditive), which encompasses all particular models studied in the literature so far.

**Definition 9.2.1** (Transportation cost). A *transportation cost* is a nondecreasing, concave, and lower semicontinuous function $\tau : [0,\infty) \to [0,\infty)$ with $\tau(0) = 0$.

**Example 9.2.1** (Branched transport, urban planning, and Steiner tree). Three particular examples of transportation costs are given by

$$\tau^{\mathrm{bt}}(m) = m^{\alpha}, \quad \tau^{\mathrm{up}}(m) = \min\{am, m+b\}, \quad \tau^{\mathrm{st}}(m) = 1 \quad \text{if } m > 0, \quad \tau^{\mathrm{st}}(0) = 0 \tag{9.9}$$

for parameters $\alpha \in (0,1)$, $a > 1$, $b > 0$. The original *branched transport* model in [38] and [23] uses $\tau^{\mathrm{bt}}$, and most analysis of transportation networks has been done for this particular case. The *urban planning* model, introduced in [6] and recast into the current framework in [8], is obtained for $\tau^{\mathrm{up}}$. Here the material sources and sinks represent the homes and workplaces of commuters, and we optimize the public transport network ($a$ has the interpretation of travel costs by other means than public transport,

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

whereas $b$ represents network maintenance costs). Finally, the *Steiner tree* problem of connecting $N$ points by a graph of minimal length can be reformulated as generalized branched transport by taking a single point as a source of mass $N-1$ and the remaining $N-1$ points as sinks of mass 1 using the transportation cost $\tau^{\mathrm{st}}$.

In the simplest formulation the generalized branched transport problem is first introduced for simple transportation networks, so-called discrete transport paths or discrete mass fluxes, which can be identified with graphs (see [9, 38]).

**Definition 9.2.2** (Discrete mass flux). Let $\mu_+ = \sum_{i=1}^{k} a_i \delta_{x_i}$ and $\mu_- = \sum_{j=1}^{l} b_i \delta_{y_j}$ be two measures with $x_i, y_j \in \mathbf{R}^n$, $a_i, b_j > 0$. Let $G$ be a weighted directed graph in $\mathbf{R}^n$ with vertices $V(G)$, edges $E(G)$, and weight function $w : E(G) \to [0, \infty)$. For an edge $e \in E(G)$, we denote by $e^+$ and $e^-$ its initial and final vertices and by $\vec{e} = \frac{e^- - e^+}{|e^- - e^+|} \in S^{n-1}$ its direction. Then the vector measure

$$\mathcal{F}_G = \sum_{e \in E(G)} w(e)(\mathcal{H}^1 \llcorner e)\vec{e} \tag{9.10}$$

is called a *discrete mass flux*. It is a discrete mass flux *between* $\mu_+$ *and* $\mu_-$ if $\operatorname{div} \mathcal{F}_G = \mu_+ - \mu_-$ in the distributional sense.

**Definition 9.2.3** (Discrete cost functional). Let $\mathcal{F}_G$ be a discrete mass flux corresponding to a graph $G$. The *discrete generalized branched transport cost functional* is given by

$$\mathcal{E}(\mathcal{F}_G) = \sum_{e \in E(G)} \tau(w(e))\mathcal{H}^1(e). \tag{9.11}$$

In the above discrete setting, the weight function $w$ encodes the amount of mass flowing through an edge, whereas the distributional divergence constraint ensures that no mass is created or lost outside the source $\mu_+$ and sink $\mu_-$ of the mass flux. Obviously, there can only be discrete mass fluxes between sources and sinks of equal mass. For general mass fluxes, described as vector-valued measures, the cost is defined via weak-* relaxation.

**Definition 9.2.4** (Continuous mass flux). Let $\mu_+, \mu_- \in \mathcal{M}_+(\mathbf{R}^n)$. A vector measure $\mathcal{F} \in \mathcal{M}(\mathbf{R}^n; \mathbf{R}^n)$ is a *(continuous) mass flux between* $\mu_+$ *and* $\mu_-$ if $\operatorname{div} \mathcal{F} = \mu_+ - \mu_-$ in the distributional sense.

**Definition 9.2.5** (Continuous cost functional). Let $\mathcal{F}$ be a continuous mass flux. The *continuous generalized branched transport cost functional* is given by

$$\mathcal{E}(\mathcal{F}) = \inf\Big\{\liminf_{k\to\infty} \mathcal{E}(\mathcal{F}_{G_k}) \;\Big|\; (\mathcal{F}_{G_k}, \operatorname{div} \mathcal{F}_{G_k}) \overset{*}{\rightharpoonup} (\mathcal{F}, \operatorname{div} \mathcal{F})\Big\}. \tag{9.12}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Th existence of minimizing mass fluxes between arbitrary prescribed sources $\mu_+$ and sinks $\mu_-$ has been shown in [9, 38] under growth conditions on the transportation cost $\tau$ near zero.

### 9.2.2 Reformulation as an image inpainting problem in 2D and convexification

In [7], we introduced a reformulation of the branched transportation energy as an image inpainting problem in two space dimensions, leading to a convexification via a functional lifting approach and to sequence (9.8) of inequalities. Here we recall the key steps of this analysis, complement it with the derivation of the opposite inequalities, and finally derive the lifted convex optimization problem, which will later form a basis of our numerical simulations.

From now on, let $\Omega \subset \mathbf{R}^2$ be open, bounded, and convex (the following could easily be generalized to Lipschitz domains which would just lead to a more technical exposition), and let $\mu_+, \mu_- \in \mathcal{M}_+(\partial\Omega)$ with equal mass $\|\mu_+\|_{\mathcal{M}} = \|\mu_-\|_{\mathcal{M}}$ denote a material source and sink supported on the boundary $\partial\Omega$. We furthermore abbreviate $V = B_1(\Omega) \subset \mathbf{R}^2$ to be the open 1-neighborhood of $\Omega$, whose sole purpose is to allow defining boundary values for images $u$ on $\Omega$ by fixing $u$ on $V \setminus \Omega$ (which is notationally easier than working with traces of BV functions).

**Remark 9.2.1** (Existence of optimal mass fluxes). In the two-dimensional setting with $\mu_+$ and $\mu_-$ concentrated on the boundary $\partial\Omega$ (which by our above assumptions on $\Omega$ satisfies $\mathcal{H}^1(\partial\Omega) < \infty$), we always have the existence of optimal (that is, $\mathcal{E}$-minimizing) mass fluxes between $\mu_+$ and $\mu_-$, independent of the choice of $\tau$. Indeed, there exists a mass flux of finite cost (for instance, a mass flux concentrated on $\partial\Omega$, which moves the mass round counterclockwise and whose cost can be bounded from above by $\tau(\|\mu_+\|_{\mathcal{M}})\mathcal{H}^1(\partial\Omega)$), so that the existence of minimizers follows from [9, Thm. 2.10].

For an image $u \in \mathrm{BV}(V)$, we can define the mass flux $\mathcal{F}_u \in \mathcal{M}(\overline{\Omega}; \mathbf{R}^2)$ as the rotated gradient of $u$,

$$\mathcal{F}_u = Du^\perp \llcorner \overline{\Omega} = \left(\nabla u^\perp \mathcal{L}^2 \llcorner V + [u] \nu_u^\perp \mathcal{H}^1 \llcorner S_u + D_c u^\perp\right) \llcorner \overline{\Omega}, \tag{9.13}$$

where $\nabla u$ denotes the approximate gradient of the image $u$, $S_u$ is the approximate discontinuity set, $\nu_u$ is the unit normal on $S_u$, $[u] = u^+ - u^-$ is the jump in function value across $S_u$ in direction $\nu_u$, $D_c u$ is a Cantor part (see, for instance, [2, § 3.9]), and $\perp$ counterclockwise rotation by $\frac{\pi}{2}$. Since $Du$ as a gradient is curl-free, $\mathcal{F}_u$ is divergence-free (in the distributional sense) in $\Omega$. It is now no surprise that fluxes between $\mu_+$ and $\mu_-$ correspond to images with particular boundary conditions. To make this correspondence explicit, let $\gamma : [0, \mathcal{H}^1(\partial\Omega)) \to \partial\Omega$ be a counterclockwise parameterization of $\partial\Omega$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

by arclength, where without loss of generality we may assume that $\gamma(0) = 0 \in \partial\Omega$ and abbreviate $\partial\Omega_t = \gamma([0,t))$.

**Definition 9.2.6** (Admissible fluxes and images). Given $\mu_+, \mu_- \in \mathcal{M}_+(\partial\Omega)$ with equal mass, we define

$$u(\mu_+,\mu_-) : V \setminus \Omega \to \mathbf{R},\ x \mapsto (\mu_+ - \mu_-)(\partial\Omega_{\gamma^{-1}(\pi_\Omega(x))}) \tag{9.14}$$

and the sets of *admissible fluxes and images* as

$$\mathcal{A}_\mathcal{F} = \{\mathcal{F} \in \mathcal{M}(\overline{\Omega};\mathbf{R}^2) \mid \mathcal{F} \text{ is mass flux between } \mu_+ \text{ and } \mu_-\}, \tag{9.15}$$

$$\mathcal{A}_u = \{u \in \mathrm{BV}(V) \mid u = u(\mu_+,\mu_-) \text{ on } V \setminus \overline{\Omega}\}. \tag{9.16}$$

By [7, Lem. 3.1.3] the mapping $u \mapsto \mathcal{F}_u \llcorner \overline{\Omega}$ from $\mathcal{A}_u$ to $\mathcal{A}_\mathcal{F}$ is bijective, so that we may also introduce the image $u_\mathcal{F} \in \mathcal{A}_u$ corresponding to the mass flux $\mathcal{F} \in \mathcal{A}_\mathcal{F}$. The relation between images and fluxes is illustrated in Figure 9.1. The following cost functional now expresses the generalized branched transport cost as a cost of images.

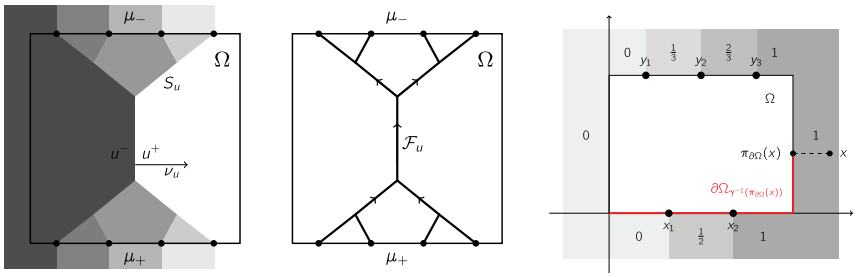



**Figure 9.1:** (a) A grey-value image $u$ and its corresponding mass flux $\mathcal{F}_u$. (b) Sketch of $u(\mu_+,\mu_-)$ for $\mu_+ = \frac{1}{2}(\delta_{x_1} + \delta_{x_2})$ and $\mu_- = \frac{1}{3}(\delta_{y_1} + \delta_{y_2} + \delta_{y_3})$, so that $u(\mu_+,\mu_-)$ takes values in $\{0, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1\}$.

**Definition 9.2.7** (Image-based cost functional). For an admissible image $u \in \mathcal{A}_u$, the *generalized branched transport cost of images* is defined as

$$\tilde{\mathcal{E}}(u) = \int_{S_u \cap \overline{\Omega}} \tau(|[u]|)\mathrm{d}\mathcal{H}^1 + \tau'(0)|Du|(\overline{\Omega} \setminus S_u), \tag{9.17}$$

where $\tau'(0) \in (0,\infty]$ denotes the right derivative of $\tau$ at 0.

In [7, Thm. 3.2.2 and Lem. 3.2.5], we proved the relation $\mathcal{E}(\mathcal{F}) \geq \tilde{\mathcal{E}}(u_\mathcal{F})$ by showing that both functionals coincide for discrete mass fluxes and the corresponding images and then by exploiting that $\tilde{\mathcal{E}}$ is lower semicontinuous, whereas $\mathcal{E}$ is the relaxation of

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

its restriction to discrete mass fluxes (that is, the largest lower semicontinuous function, which coincides with $\tilde{\mathcal{E}}$ on discrete mass fluxes). The opposite inequality can be obtained by showing that $\tilde{\mathcal{E}}$ is a relaxation as well, an issue which was considered in [9, 24].

**Theorem 9.2.1** (Equality of flux-based and image-based cost). *Let $\mu_+, \mu_- \in \mathcal{M}_+(\partial\Omega)$ with equal mass. For a mass flux $\mathcal{F} \in \mathcal{A}_{\mathcal{F}}$ and the corresponding image $u_{\mathcal{F}} \in \mathcal{A}_u$, we have $\mathcal{E}(\mathcal{F}) = \tilde{\mathcal{E}}(u_{\mathcal{F}})$.*

*Proof.* Since the relation between $\mathcal{A}_u$ and $\mathcal{A}_{\mathcal{F}}$ is one-to-one, it suffices to show that $\mathcal{E}(\mathcal{F}_u) = \tilde{\mathcal{E}}(u)$ for any $u \in \mathcal{A}_u$. Now note that $\mathcal{F}_u = \theta(\mathcal{H}^1 \llcorner S) + \mathcal{F}^d$ for $S = S_u \cap \overline{\Omega}$, $\theta = [u]\nu_u^\perp$, and $\mathcal{F}^d = (\nabla u^\perp \mathcal{L}^2 \llcorner V + D_c u^\perp) \llcorner \overline{\Omega}$. By [2, Thm. 3.78] $S$ is countably $\mathcal{H}^1$-rectifiable, and by [2, Lem. 3.76] $\mathcal{F}^d$ is diffuse, that is, singular with respect to $\mathcal{H}^1 \llcorner R$ for any countably 1-rectifiable $R \subset \overline{\Omega}$. Thus by [9, Prop. 2.32] we have

$$\mathcal{E}(\mathcal{F}_u) = \int_S \tau(|\theta|)\,\mathrm{d}\mathcal{H}^1 + \tau'(0)|\mathcal{F}^d|(\overline{\Omega}); \tag{9.18}$$

however, this equals exactly $\tilde{\mathcal{E}}(u)$. □

It turns out that $\tilde{\mathcal{E}}$ can be expressed as an energy of a surface in $\mathbf{R}^3$, which will lead to a convex optimization problem. This approach has been introduced in [1] to prove optimality of special solutions to the Mumford–Shah problem (and related ones) by exhibiting a lower bound on the surface energy, and it was subsequently exploited in [31, 32] to numerically compute global minimizers of the Mumford–Shah functional. The setting in [1, 31, 32] is slightly more general than we need here. The authors consider a generalized Mumford–Shah functional

$$J(u) = \int_V g(x, u(x), \nabla u(x))\,\mathrm{d}x + \int_{S_u} h(x, u^+, u^-, \nu_u)\,\mathrm{d}\mathcal{H}^1(x), \tag{9.19}$$

where $g$ is a normal Carathéodory function convex in its third argument, and $h$ is one-homogeneous and convex in its last argument and subadditive in $(u^+, u^-)$ (see [2, § 5.2–5.3] for details on the requirements). In [1, 32], it is shown that $J(u)$ can be estimated from below as follows. Let

$$1_u : V \times \mathbf{R} \to \{0,1\}, \quad 1_u(x,s) = \begin{cases} 1 & \text{if } u(x) > s, \\ 0 & \text{otherwise} \end{cases} \tag{9.20}$$

denote the characteristic function of the subgraph of the image $u \in \mathrm{BV}(V)$ and introduce the convex set

$$\mathcal{K} = \left\{ \phi = (\phi^x, \phi^s) \in C_0^\infty(V \times \mathbf{R}; \mathbf{R}^2 \times \mathbf{R}) \;\middle|\; \phi^s(x,s) \geq g^*\big(x, s, \phi^x(x,s)\big)\ \forall\, (x,s) \in V \times \mathbf{R}, \right.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$$\left| \int_{s_1}^{s_2} \phi^x(x,s)\mathrm{d}s \right| \leq h(x,s_1,s_2,\nu) \; \forall\, x \in V, s_1 < s_2, \nu \in S^1 \Bigg\} \tag{9.21}$$

of three-dimensional vector fields, where $g^*$ denotes the Legendre–Fenchel conjugate of $g$ with respect to its last argument. Then the generalized Mumford–Shah functional can be estimated via

$$J(u) \geq \sup_{\phi \in \mathcal{K}} \int_{V \times \mathbf{R}} \phi \cdot \mathrm{d}D1_u, \tag{9.22}$$

where the right-hand integral can be interpreted as an integral over the complete graph of $u$ and thus as a surface functional. Even equality is expected but has not been rigorously proved. The above can be specialized to our setting by picking $g(x,u,p) = \tau'(0)|p|$ and $h(x,u^+,u^+,\nu) = \tau(|u^+ - u^-|)$ for which $g^*(x,u,q) = 0$ if $|q| \leq \tau'(0)$ and $g^*(x,u,q) = \infty$ else (so that the condition $\phi^s(x,s) \geq g^*(x,s,\phi^x(x,s))$ turns into the inequalities $|\phi^x(x,s)| \leq \tau'(0)$ and $\phi^s(x,s) \geq 0$).

**Definition 9.2.8** (Surface-based cost functional). Let $\mu_+, \mu_- \in \mathcal{M}_+(\partial\Omega)$ with equal mass. We set

$$\begin{aligned} \mathcal{K} = \Bigg\{ & \phi = (\phi^x, \phi^s) \in C_0^\infty(V \times \mathbf{R}; \mathbf{R}^2 \times \mathbf{R}) \;\Bigg| \\ & |\phi^x(x,s)| \leq \tau'(0),\ \phi^s(x,s) \geq 0 \; \forall\, (x,s) \in V \times \mathbf{R}, \\ & \left| \int_{s_1}^{s_2} \phi^x(x,s)\mathrm{d}s \right| \leq \tau(s_2 - s_1) \; \forall\, x \in V, s_1 < s_2 \Bigg\}. \end{aligned} \tag{9.23}$$

For an admissible image $u \in \mathcal{A}_u$, the *generalized branched transport cost of surfaces* is defined as

$$\mathcal{G}(1_u) = \sup_{\phi \in \mathcal{K}} \int_{\overline{\Omega} \times \mathbf{R}} \phi \cdot \mathrm{d}D1_u. \tag{9.24}$$

In [1, 7], it is shown that $\tilde{\mathcal{E}}(u) \geq \mathcal{G}(1_u)$. We now show equality.

**Theorem 9.2.2** (Equality of image-based and surface-based cost). *Let $\mu_+, \mu_- \in \mathcal{M}_+(\partial\Omega)$ with equal mass and assume without loss of generality that $u(\mu_+, \mu_-)$ takes the minimum value 0 and maximum value $M \leq \|\mu_+\|_{\mathcal{M}}$. For an image $u \in \mathcal{A}_u$, we have $\tilde{\mathcal{E}}(u) = \mathcal{G}(1_u)$. Moreover,*

$$\min_{u \in \mathcal{A}_u} \mathcal{G}(1_u) = \min_{u \in \mathcal{A}_u \cap \mathrm{BV}(V;[0,M])} \mathcal{G}(1_u) = \min_{u \in \mathcal{A}_u \cap \mathrm{BV}(V;[0,M])} \sup_{\phi \in \tilde{\mathcal{K}}} \int_{\overline{\Omega} \times [0,M]} \phi \cdot \mathrm{d}D1_u \tag{9.25}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*for the set*

$$\tilde{\mathcal{K}} = \Big\{ \phi = (\phi^x, \phi^s) \in C^0(\overline{\Omega} \times [0,M]; \mathbf{R}^2 \times [0,\infty)) \;\Big|\; |\phi^x(x,s)| \le \tau'(0),\ \phi^s(x,s) \ge 0 \ \forall\, (x,s) \in \overline{\Omega} \times [0,M], \left| \int_{s_1}^{s_2} \phi^x(x,s)\mathrm{d}s \right| \le \tau(s_2 - s_1) \ \forall\, x \in \overline{\Omega}, 0 \le s_1 < s_2 \le M \Big\}. \tag{9.26}$$

*Proof.* We need to show that $\tilde{\mathcal{E}}(u) \le \mathcal{G}(1_u)$. To this end, it suffices to consider $\tau$ with $\tau(m) = \alpha m$ for all $m \le m_0$, where $\alpha < \infty$ and $m_0 > 0$ are arbitrary. Indeed, assume equality for such transportation costs, let $\tau$ be a given transportation cost, and set $\tau_n(m) = \min\{\tau(m), \alpha_n m\}$ for a sequence $0 < \alpha_1 < \alpha_2 < \cdots$ with $\alpha_n \to \tau'(0)$ as $n \to \infty$. Decorating $\mathcal{G}$ and $\tilde{\mathcal{E}}$ with a superscript to indicate what transportation cost they are based on, we have

$$\mathcal{G}^\tau(1_u) \ge \mathcal{G}^{\tau_n}(1_u) = \tilde{\mathcal{E}}^{\tau_n}(u) \to \tilde{\mathcal{E}}^\tau(u) \quad \text{as } n \to \infty \tag{9.27}$$

by monotone convergence, as desired.

By [2, Thm. 3.78] $S_u$ is countably $\mathcal{H}^1$-rectifiable. Furthermore, $[u] \in L^1(\mathcal{H}^1 \llcorner S_u)$. Thus the jump part of $Du$ can be treated via a decomposition strategy as, for instance, also used in [13, Lem. 4.2]. In detail, let $\varepsilon > 0$ be arbitrary. Since $S_u$ is rectifiable, there exists a compact (oriented) $C^1$-manifold $\mathcal{N} \subset \mathbf{R}^2$ with $|Du|(S_u \setminus \mathcal{N}) < \varepsilon$. For $x \in \mathcal{N}$ and $\ell > 0$, let us denote by $F_x^\ell \subset \mathbf{R}^2$ the closed square of side length $2\ell$, centered at $x$ and axis-aligned with the tangent and the normal vector to $\mathcal{N}$ in $x$. Also, denote by $R_x : \mathbf{R}^2 \to \mathbf{R}^2$ the rigid motion that maps $x$ to 0 and the unit tangent of $\mathcal{N}$ in $x$ to $(0,1)$ (thus $R_x(F_x^\ell) = [-\ell,\ell]^2$). Now fix $\delta > 0$ such that

$$R_x(\mathcal{N} \cap F_x^\delta) \text{ is the graph of a map } g_x \in C^1([-\delta,\delta];[-\delta,\delta]) \text{ with } g_x(0) = g_x'(0) = 0, \tag{9.28}$$

$$|Du|\Big( \bigcup_{x \in \mathcal{N}} F_x^\delta \setminus S_u \Big) < \varepsilon \tag{9.29}$$

(the latter can be achieved since $\bigcup_{x\in\mathcal{N}} F_x^\delta \setminus S_u \to \mathcal{N} \setminus S_u$ monotonically as $\delta \to 0$, and thus by outer regularity of $|Du|$ we have $|Du|(\bigcup_{x\in\mathcal{N}} F_x^\delta \setminus S_u) \to |Du|(\mathcal{N} \setminus S_u) = 0$ as $\delta \to 0$). Now $\mathcal{N} \subset \bigcup_{x\in\mathcal{N}} \bigcup_{\ell<\delta} F_x^\ell$, so that by the Vitali–Besicovitch covering theorem [2, Thm. 2.19] there is a countable disjoint family of cubes $F_{x_1}^{\ell_1}, F_{x_2}^{\ell_2}, \ldots, x_i \in \mathcal{N}$, $\ell_i < \delta$, whose union $F$ satisfies $|Du|(\mathcal{N} \setminus F) = 0$. By taking a finite subfamily $F_{x_1}^{\ell_1}, \ldots, F_{x_K}^{\ell_K}$ we achieve $|Du|(\mathcal{N} \setminus \bigcup_{k=1}^K F_{x_k}^{\ell_k}) < \varepsilon$.

On $F_{x_i}^{\ell_i}$ define the projection $p_i : F_{x_i}^{\ell_i} \to \mathcal{N}$ by

$$p_i = R_{x_i}^{-1} \circ g_{x_i} \circ \pi_{\mathbf{R}\times\{0\}} \circ R_{x_i} \tag{9.30}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

(where $\pi_{\mathbf{R}\times\{0\}}$ returns first coordinate of a vector; $p_i$ is the projection along one axis direction of $F_{x_i}^{\ell_i}$). Furthermore, define $\psi$ as

$$\tilde{\psi}_i(x,s) = \begin{cases} \nu_{\mathcal{N}}(p_i(x))\frac{\tau(|[u](p_i(x))|)}{[u](p_i(x))} \\ \quad \text{if } \operatorname{dist}(x,\partial F_{x_i}^{\ell_i}) > \eta \text{ and} \\ \quad s \in [\min\{u^-(p_i(x)), u^+(p_i(x))\}, \max\{u^-(p_i(x)), u^+(p_i(x))\}], \\ 0 \quad \text{else,} \end{cases} \tag{9.31}$$

with $\nu_{\mathcal{N}}$ being the normal vector to $\mathcal{N}$. Note that by construction we have $|\int_{s_1}^{s_2} \tilde{\psi}_i(x,s)\,\mathrm{d}s| \le \tau(s_2 - s_1)$ for all $s_1 < s_2$ and $x \in F_{x_i}^{\ell_i}$, as well as $|\tilde{\psi}_i(x,s)| \le \alpha$ due to $\tau(m) \le \alpha m$. Mollifying $\tilde{\psi}_i$ with mollifier $\rho_\eta(x) = \rho(x/\eta)/\eta$ for $\rho \in C_0^\infty([-1,1]^3;[0,\infty))$ with unit integral, the above constraints stay satisfied by Jensen's inequality, and we obtain some $\psi_i = \rho_\eta * \tilde{\psi}_i \in C_0^\infty(F_{x_i}^{\ell_i} \times \mathbf{R}; \mathbf{R}^2)$. Extending $\psi_i$ by zero to $V \times \mathbf{R}$, we can define $\phi_i \in \mathcal{K}$ as $\phi_i(x,s) = (\psi_i(x,s), 0)$. We now set $\hat{\phi} = \sum_{i=1}^K \phi_i$. Note that we can choose $\eta$ small enough such that $\int_{\overline{\Omega}\times\mathbf{R}} \hat{\phi} \cdot \mathrm{d}D1_u \ge \int_{S_u} \tau(|[u]|)\,\mathrm{d}\mathcal{H}^1 - 5\alpha\varepsilon$. Indeed, abbreviating

$$\chi_{(a,b)}(s) = \begin{cases} 1 & \text{if } a < s < b, \\ -1 & \text{if } b < s < a, \\ 0 & \text{else,} \end{cases} \tag{9.32}$$

we can calculate

$$\begin{aligned} \int_{\overline{\Omega}\times\mathbf{R}} \hat{\phi} \cdot \mathrm{d}D1_u &= \int_{(\overline{\Omega}\cap\bigcup_{i=1}^K F_{x_i}^{\ell_i})\times\mathbf{R}} \hat{\phi} \cdot \mathrm{d}D1_u \ge \int_{(\overline{\Omega}\cap S_u\cap\bigcup_{i=1}^K F_{x_i}^{\ell_i})\times\mathbf{R}} \hat{\phi} \cdot \mathrm{d}D1_u - \alpha\varepsilon \\ &= \sum_{i=1}^K \int_{\overline{\Omega}\cap\mathcal{N}\cap F_{x_i}^{\ell_i}} \int_{\mathbf{R}} \chi_{(u^-(x),u^+(x))}(s)\nu_{\mathcal{N}}(x) \cdot \psi_i(x,s)\,\mathrm{d}s\,\mathrm{d}\mathcal{H}^1(x) - 2\alpha\varepsilon \\ &\xrightarrow[\eta\to 0]{} \sum_{i=1}^K \int_{\overline{\Omega}\cap\mathcal{N}\cap F_{x_i}^{\ell_i}} \int_{\mathbf{R}} \chi_{(u^-(x),u^+(x))}(s)\nu_{\mathcal{N}}(x) \cdot \tilde{\psi}_i(x,s)\,\mathrm{d}s\,\mathrm{d}\mathcal{H}^1(x) - 2\alpha\varepsilon \\ &= \sum_{i=1}^K \int_{\overline{\Omega}\cap\mathcal{N}\cap F_{x_i}^{\ell_i}} \tau\big(\big|[u](x)\big|\big)\,\mathrm{d}\mathcal{H}^1(x) - 2\alpha\varepsilon \ge \int_{\overline{\Omega}\cap\mathcal{N}} \tau\big(\big|[u](x)\big|\big)\,\mathrm{d}\mathcal{H}^1(x) - 3\alpha\varepsilon \\ &\ge \int_{\overline{\Omega}\cap S_u} \tau\big(\big|[u](x)\big|\big)\,\mathrm{d}\mathcal{H}^1(x) - 4\alpha\varepsilon \end{aligned} \tag{9.33}$$

since as $\eta \to 0$, $\psi_i$ converge to $\tilde{\psi}_i$ in $L^1(\mathcal{H}^2 \llcorner \mathcal{N} \times \mathbf{R})$, and $(x,s) \mapsto \chi_{(u^-(x),u^+(x))}(s)\nu_{\mathcal{N}}(x)$ is in $L^\infty(\mathcal{H}^2 \llcorner \mathcal{N} \times \mathbf{R})$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Now consider the cost associated with the diffuse part of $Du$. By [2, Prop. 3.64] $u_\xi = \rho_\xi * u \to \tilde{u}$ pointwise on $V \setminus S_u$, where $\rho_\xi$ is some mollifier with length scale $\xi$, and $\tilde{u}$ is a particular representative of $u$, the so-called approximate limit. Consequently, $u_\xi \to \tilde{u}$ pointwise $|Du| \llcorner V \setminus S_u$-almost everywhere. Thus by Egorov's theorem there exists a measurable set $B \subset V$ such that $|Du|((V \setminus S_u) \setminus B) < \varepsilon$ and $u_\xi \to \tilde{u}$ uniformly on $B$. Let $\xi$ be small enough such that $|\tilde{u} - u_\xi| < m_0/4$ on $B$, and let $\psi \in C_0^\infty(V; \mathbf{R}^2)$ be such that $|\psi| \le 1$ everywhere and $\int_{\overline{\Omega}} \psi \cdot \mathrm{d}Du \ge |Du|(\overline{\Omega}) - \varepsilon$. Furthermore, fix $\eta > 0$ such that $|Du|(U_\eta \setminus \bigcup_{k=1}^K F_{x_k}^{\ell_k}) < \varepsilon$ for the $\eta$-neighborhood $U_\eta$ of $\partial V \cup \bigcup_{k=1}^K F_{x_k}^{\ell_k}$. We now define

$$\bar{\phi}(x,s) = \begin{pmatrix} \alpha\psi(x)\chi_1(x)\chi_2(t - u_\xi(x)) \\ 0 \end{pmatrix}, \tag{9.34}$$

where $\chi_1 \in C_0^\infty(V; [0,1])$ is a cutoff function that is zero on $\bigcup_{k=1}^K F_{x_k}^{\ell_k}$ and one outside $U_\eta$, and where $\chi_2 \in C_0^\infty(\mathbf{R}; [0,1])$ is a cutoff function that is one on $[-m_0/4, m_0/4]$ and zero outside $[-m_0/2, m_0/2]$. Note that $\bar{\phi} \in \mathcal{K}$ by construction and

$$\begin{aligned}
\int_{\overline{\Omega}\times\mathbf{R}} \bar{\phi} \cdot \mathrm{d}D1_u &= \alpha \int_{\{(x,\tilde{u}(x))|x\in\overline{\Omega}\}} \chi_1(x) \begin{pmatrix} \psi(x) \\ 0 \end{pmatrix} \cdot \mathrm{d}D1_u(x,s) \\
&\ge \alpha \int_{\{(x,\tilde{u}(x))|x\in\overline{\Omega}\setminus S_u\}} \begin{pmatrix} \psi \\ 0 \end{pmatrix} \cdot \mathrm{d}D1_u - \alpha|Du|(S_u \setminus U_\eta) \\
&\quad - \alpha|Du|\left( U_\eta \setminus \bigcup_{k=1}^K F_{x_k}^{\ell_k} \right) - \alpha|Du|((\overline{\Omega} \setminus S_u) \setminus B) \\
&\ge \alpha \int_{\{(x,\tilde{u}(x))|x\in\overline{\Omega}\setminus S_u\}} \begin{pmatrix} \psi \\ 0 \end{pmatrix} \cdot \mathrm{d}D1_u - 3\alpha\varepsilon = \alpha \int_{\mathbf{R}} \int_{\overline{\Omega}\setminus S_u} \psi \cdot \mathrm{d}D\chi_{\{u>s\}}\,\mathrm{d}s - 3\alpha\varepsilon \\
&= \alpha \int_{\overline{\Omega}\setminus S_u} \psi \cdot \mathrm{d}Du - 3\alpha\varepsilon \ge \alpha|Du|(\overline{\Omega} \setminus S_u) - 4\alpha\varepsilon,
\end{aligned} \tag{9.35}$$

where $\chi_{\{u>s\}}$ is the characteristic function of the $s$-superlevel set of $u$, and where in the last equality we used the coarea formula.

Summarizing, we have $\phi = \hat{\phi} + \bar{\phi} \in \mathcal{K}$ with

$$\int_{\overline{\Omega}\times\mathbf{R}} \phi \cdot \mathrm{d}D1_u \ge \int_{\overline{\Omega}\cap S_u} \tau\big(\big|[u](x)\big|\big)\,\mathrm{d}\mathcal{H}^1(x) + \alpha|Du|(\overline{\Omega} \setminus S_u) - 8\alpha\varepsilon = \tilde{\mathcal{E}}(u) - 8\alpha\varepsilon,$$

and thus $\mathcal{G}(1_u) \ge \tilde{\mathcal{E}}(u)$ follows from the arbitrariness of $\varepsilon$.

From the definition of $\tilde{\mathcal{E}}$ it is obvious that $\tilde{\mathcal{E}}(u)$ decreases if $u$ is clipped to the range $[0, M]$. Thus minimizers of $\mathcal{G}(1_u) = \tilde{\mathcal{E}}(u)$ among all admissible images $u$ lie in $\mathcal{A}_u \cap \mathrm{BV}(V; [0, M])$, and we may restrict the integral in the definition of $\mathcal{G}$ to $\overline{\Omega} \times [0, M]$. Finally, by density of $\{\phi \in \mathcal{K} \mid \phi^s = 0\}$ in $\tilde{\mathcal{K}}$ with respect to the supremum norm we may replace $\mathcal{K}$ with $\tilde{\mathcal{K}}$ without changing the supremum. □

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Note that we could even set $\phi^s \equiv 0$ in $\tilde{\mathcal{K}}$ without changing $\sup_{\phi\in\tilde{\mathcal{K}}} \int_{\overline{\Omega}\times[0,M]} \phi \cdot \mathrm{d}D1_u$ since the integral increases if $\phi^s$ decreases. The problem of minimizing $\mathcal{G}(1_u)$ among all characteristic functions of subgraphs of admissible images is not convex, since the space of characteristic functions is not. The underlying idea of [1, 7, 31] is that we do not lose much by convexifying the domain of $\mathcal{G}$ as follows.

**Definition 9.2.9** (Convex cost functional). Let $\mu_+, \mu_- \in \mathcal{M}_+(\partial\Omega)$ with equal mass, and let $M$, $\tilde{\mathcal{K}}$ be as in Theorem 9.2.2. We set

$$\mathcal{C} = \{v \in \mathrm{BV}(V \times \mathbf{R}; [0,1]) \mid v = 1_{u(\mu_+,\mu_-)} \text{ on } (V \times \mathbf{R}) \setminus (\overline{\Omega} \times [0,M])\}, \tag{9.36}$$

where we extended $1_{u(\mu_+,\mu_-)}$ by 1 to $V \times (-\infty, 0)$ and by 0 to $V \times (M, \infty)$. The *convex generalized branched transport cost* is $\tilde{\mathcal{G}} : \mathcal{C} \to \mathbf{R}$ defined as

$$\tilde{\mathcal{G}}(v) = \sup_{\phi\in\tilde{\mathcal{K}}} \int_{\overline{\Omega}\times[0,M]} \phi \cdot \mathrm{d}Dv. \tag{9.37}$$

By definition and Theorem 9.2.2, $\tilde{\mathcal{G}}$ coincides with $\mathcal{G}$ on functions of the form $v = 1_u$ with $u \in \mathcal{A}_u$. The following proposition shows that the problem of minimizing $\tilde{\mathcal{G}}$ is related to the original generalized branched transport problem in the sense that if the minimizer of $\tilde{\mathcal{G}}$ is binary (that is, if it only takes the values 0 and 1), then it is a solution of the original problem. The proposition also shows that the original and the convex minimization problem cannot be fully equivalent since sometimes $\tilde{\mathcal{G}}$ has nonbinary minimizers (however, those nonbinary minimizers may coexist with binary minimizers so that the minimization problems might still be equivalent after selecting the binary minimizers).

**Proposition 9.2.1** (Properties of convex cost functional). *Let $\mu_+, \mu_- \in \mathcal{M}_+(\partial\Omega)$ with equal mass.*

1. *$\tilde{\mathcal{G}}$ is convex weakly-* lower semicontinuous and satisfies $\inf_{v\in\mathcal{C}} \tilde{\mathcal{G}}(v) \leq \min_{u\in\mathcal{A}_u} \mathcal{G}(1_u)$.*
2. *If a minimizer $v \in \mathcal{C}$ of $\tilde{\mathcal{G}}$ is binary, then $v = 1_u$ for a minimizer $u \in \mathcal{A}_u$ of $\mathcal{G}(1_u)$.*
3. *If $\tau$ is not linear, then there exist $\mu_+, \mu_- \in \mathcal{M}_+(\partial\Omega)$ such that if $\tilde{\mathcal{G}}$ has minimizers, then at least some of them are nonbinary.*

*Proof.*

1. As the supremum over linear functionals on a convex domain, $\tilde{\mathcal{G}}$ is convex and lower semicontinuous with respect to the weak-* topology. Furthermore, $\inf_v \tilde{\mathcal{G}}(v) \leq \inf_{u\in\mathcal{A}_u} \tilde{\mathcal{G}}(1_u) = \min_{u\in\mathcal{A}_u} \mathcal{G}(1_u)$.
2. First, note that $\tilde{\mathcal{G}}(v) = \infty$ unless $v$ is monotonically decreasing in the $s$-direction. Indeed, if $(Dv)_3$ is not nonpositive, then there exists a continuous $\phi^s \geq 0$ with $\int_{\overline{\Omega}\times[0,M]} \phi^s \, \mathrm{d}(Dv)_3 > 0$ (for instance, take the positive part of some $\psi \in C^0(\overline{\Omega} \times [0,M])$ with $\int_{\overline{\Omega}\times[0,M]} \psi \, \mathrm{d}(Dv)_3 \approx \|(Dv)_3\|_{\mathcal{M}}$) such that $\tilde{\mathcal{G}}(v) \geq \sup_{\lambda>0} \int_{\overline{\Omega}\times[0,M]} (0, 0,$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

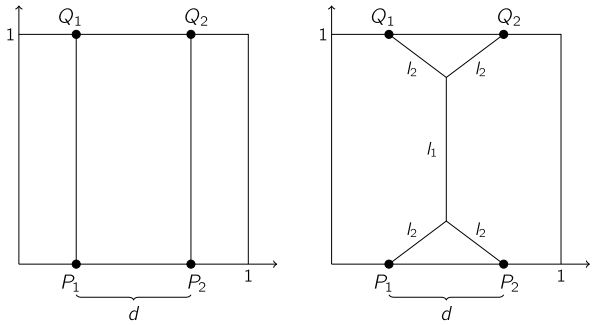



**Figure 9.2:** Generalized branched transport in $\Omega = [0,1]^2$ from two points at $P_1, P_2 \in \partial\Omega$ with equal mass to two points $Q_1, Q_1 \in \partial\Omega$ with equal mass. Depending on the mass and the point distance $d$, the optimal network has the left or right topology. At the bifurcation point, both topologies are optimal.

$\lambda\phi^s) \cdot \mathrm{d}Dv = \infty$. Thus $v$ can be represented as $1_u$ for some function $u \in \mathcal{A}_u$. Due to the previous point, $1_u$ must be a minimizer of $\mathcal{G}$.

3. Assume the contrary, that is, for any $\mu_+, \mu_- \in \mathcal{M}_+(\partial\Omega)$ with equal mass, the minimizers of $\tilde{\mathcal{G}}$ are binary. Since $\tau$ is not linear, there exist $\mu_+$, $\mu_-$ such that the corresponding generalized branched transport problem has no unique minimizer (see, for instance, Figure 9.2). Thus there are $u_1, u_2 \in \mathcal{A}_u$, $u_1 \neq u_2$, with $\min_{u\in\mathcal{A}_u} \mathcal{G}(1_u) = \mathcal{G}(1_{u_1}) = \mathcal{G}(1_{u_2}) = \tilde{\mathcal{G}}(1_{u_1}) = \tilde{\mathcal{G}}(1_{u_2}) = \min_{v\in\mathcal{C}} \tilde{\mathcal{G}}(v)$, where the last equality follows from the previous point. However, since $\mathcal{C}$ and $\tilde{\mathcal{G}}$ are convex, $(1_{u_1} + 1_{u_2})/2$ is also a minimizer of $\tilde{\mathcal{G}}$, which is nonbinary. □

When surface energies are relaxed to energies over functions $v \in \mathrm{BV}(V \times \mathbf{R}; [0,1])$ as in our case, the coarea formula is typically used to show that for a minimizer $v$, the characteristic functions of its superlevel sets have the same minimizing cost, and thus there are always binary minimizers. In the case of a one-homogeneous $\tau(m) = \alpha m$, this works as follows:

$$\tilde{\mathcal{G}}(v) = \int_{\overline{\Omega}\times[0,M]} \tau\left(\frac{Dv}{|Dv|}\right) \mathrm{d}Dv = \int_0^1 \int_{\overline{\Omega}\times[0,M]} \tau\left(\frac{D\chi_{\{v>t\}}}{|D\chi_{\{v>t\}}|}\right) \mathrm{d}D\chi_{\{v>t\}}\,\mathrm{d}t = \int_0^1 \tilde{\mathcal{G}}(\chi_{\{v>t\}})\,\mathrm{d}t, \quad (9.38)$$

where we exploited the one-homogeneity of $\tau$ and the coarea formula (a similar calculation can be performed for the lifting of the generalized Mumford–Shah functional $J$ with $h \equiv 0$). If $v$ is a minimizer so that $\tilde{\mathcal{G}}(v) \leq \tilde{\mathcal{G}}(\chi_{\{v>t\}})$ for all $t$, then by the above equality we necessarily have $\tilde{\mathcal{G}}(v) = \tilde{\mathcal{G}}(\chi_{\{v>t\}})$ for almost all $t \in [0,1]$. However, a formula as the above is not true in our case.

**Proposition 9.2.2** (Convex cost of superlevel sets). *We have* $\tilde{\mathcal{G}}(v) \leq \int_0^1 \tilde{\mathcal{G}}(\chi_{\{v>t\}})\,\mathrm{d}t$, *where depending on* $\tau$ *and* $v$ *the inequality may be strict.*

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*Proof.* The inequality holds by the convexity of $\tilde{\mathcal{G}}$ and Jensen's inequality in combination with $v = \int_0^1 \chi_{\{v>t\}}\,\mathrm{d}t$. To show that the inequality is sometimes strict, first, note that by an analogous construction as in the proof of Theorem 9.2.2 we have $\tilde{\mathcal{G}}(v) = \tilde{\mathcal{G}}_1(v) + \tilde{\mathcal{G}}_2(v) = \sup_{\phi\in\tilde{\mathcal{K}}} \int_{\partial\Omega\times\mathbf{R}} \phi\cdot \mathrm{d}Dv + \sup_{\phi\in\tilde{\mathcal{K}}} \int_{\Omega\times\mathbf{R}} \phi\cdot \mathrm{d}Dv$ with $\tilde{\mathcal{G}}_i(v) \le \int_0^1 \tilde{\mathcal{G}}_i(\chi_{\{v>t\}})\,\mathrm{d}t$, $i = 1, 2$, for the same reason as above. Now consider the example $v(x,s) = \int_0^1 1_{u+tc}(x,s)\,\mathrm{d}t$ for $(x,s) \in \Omega\times\mathbf{R}$ with $u(x) = mx_1$ for some $c, m > 0$. We have

$$\tilde{\mathcal{G}}_2(\chi_{\{v>t\}}) = \mathcal{G}(1_{u+tc}) = \tilde{\mathcal{E}}(u) = \tau'(0)m\mathcal{L}^2(\Omega), \tag{9.39}$$

whereas

$$\begin{aligned}
\tilde{\mathcal{G}}_2(v) &\le \int_\Omega \sup_{\phi\in\tilde{\mathcal{K}}} \int_0^M \phi(x,s)\cdot\nabla v(x,s)\,\mathrm{d}s\,\mathrm{d}x = \int_\Omega \sup_{\phi\in\tilde{\mathcal{K}}} \int_{u(x)-c}^{u(x)} \phi_1(x,s)\frac{m}{c}\,\mathrm{d}s\,\mathrm{d}x \\
&= \mathcal{L}^2(\Omega)\frac{m}{c}\sup\Bigg\{\int_0^c \psi\,\mathrm{d}s \;\Bigg|\; \psi : [0,c]\to\mathbf{R}, \\
&\qquad \left|\int_{s_1}^{s_2} \psi(s)\,\mathrm{d}s\right| \le \tau(s_2-s_1) \text{ for all } 0\le s_1 < s_2 \le c\Bigg\} \\
&= \mathcal{L}^2(\Omega)m\frac{\tau(c)}{c}.
\end{aligned} \tag{9.40}$$

Summarizing, $\tilde{\mathcal{G}}_2(v) \le \mathcal{L}^2(\Omega)m\frac{\tau(c)}{c} < \tau'(0)m\mathcal{L}^2(\Omega) = \int_0^1 \tilde{\mathcal{G}}_2(\chi_{\{v>t\}})\,\mathrm{d}t$, as desired. □

This does not imply that $\tilde{\mathcal{G}}$ does not always have binary minimizers; intuitively, since nonbinary functions $v$ may have smaller costs in the domain interior, we have to pay some extra cost for the transition from binary on $\partial\Omega\times[0,M]$ to nonbinary in $\Omega\times[0,M]$. The next section and the numerical experiments provide evidence that binary minimizers exist at least in many relevant cases, as is also believed for the generalized Mumford–Shah setting.

The last remaining inequality in (9.8) bounds the convex saddle point problem $\inf_{v\in\mathcal{C}} \sup_{\phi\in\tilde{\mathcal{K}}} \int_{\overline{\Omega}\times[0,M]} \phi\cdot\mathrm{d}Dv$ below by the corresponding primal optimization problem in the vector field $\phi$ (note that in (9.8), we did not reduce $\mathcal{K}$ to $\tilde{\mathcal{K}}$ for simplicity of exposition). To have an equality, we thus need to show strong duality.

**Theorem 9.2.3** (Strong duality for convex cost). *Let $\mu_+$, $\mu_-$, $\mathcal{C}$, $M$, $\tilde{\mathcal{K}}$ be as in Definition 9.2.9. Then $\tilde{\mathcal{G}}$ has a minimizer, and we have the strong duality*

$$\begin{aligned}
\min_{v\in\mathcal{C}} \tilde{\mathcal{G}}(v) &= \min_{v\in\mathcal{C}} \sup_{\phi\in\tilde{\mathcal{K}}} \int_{\overline{\Omega}\times[0,M]} \phi\cdot\mathrm{d}Dv = \sup_{\phi\in\tilde{\mathcal{K}}} \inf_{v\in\mathcal{C}} \int_{\overline{\Omega}\times[0,M]} \phi\cdot\mathrm{d}Dv \\
&= \sup_{\phi\in\tilde{\mathcal{K}}\cap\mathcal{C}^1(\overline{\Omega}\times[0,M];\mathbf{R}^2\times\mathbf{R})} \mathcal{D}(\phi)
\end{aligned} \tag{9.41}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$$\text{for } \mathcal{D}(\phi) = \int_{\partial(\Omega\times(0,M))} 1_{u(\mu_+,\mu_-)}\phi\cdot n\,\mathrm{d}\mathcal{H}^2 - \int_{\Omega\times(0,M)} \max\{0,\operatorname{div}\phi\}\,\mathrm{d}x\,\mathrm{d}s. \tag{9.42}$$

*Proof.* The last equality is obtained via the integration by parts

$$\int_{\overline{\Omega}\times[0,M]} \phi\cdot \mathrm{d}Dv = \int_{\partial(\Omega\times(0,M))} v\phi\cdot n\,\mathrm{d}\mathcal{H}^2 - \int_{\Omega\times(0,M)} v\operatorname{div}\phi\,\mathrm{d}x\,\mathrm{d}s, \tag{9.43}$$

noticing that $v = 1_{u(\mu_+,\mu_-)}$ on $\partial\Omega\times[0,M]$ and taking in $\Omega\times(0,M)$ the maximizing $v = 1$ if $\operatorname{div}\phi > 0$ and 0 else (we also exploited the denseness of $C^1$ in $C^0$). As for the first equality, the strong duality, we define $\mathcal{X} = C^1(\overline{\Omega}\times[0,M];\mathbf{R}^2\times\mathbf{R})$, $\mathcal{Y} = C^0(\overline{\Omega}\times[0,M])\times C^0(\overline{\Omega}\times[0,M];\mathbf{R}^2\times\mathbf{R})$, and

$$A:\mathcal{X}\to\mathcal{Y},\quad A\phi = (\operatorname{div}\phi,\phi), \tag{9.44}$$

$$F:\mathcal{X}\to[0,\infty],\quad F(\phi) = \iota_{\tilde{\mathcal{K}}}(\phi), \tag{9.45}$$

$$G:\mathcal{Y}\to\mathbf{R},\quad G(\psi_1,\psi_2) = \int_{\Omega\times(0,M)} \max\{0,\psi_1\}\,\mathrm{d}x\,\mathrm{d}s - \int_{\partial(\Omega\times(0,M))} 1_{u(\mu_+,\mu_-)}\psi_2\cdot n\,\mathrm{d}\mathcal{H}^2. \tag{9.46}$$

The mapping $A$ is bounded linear, whereas $F$ and $G$ are proper convex lower semicontinuous. Furthermore, $0\in\operatorname{int}(\operatorname{dom}G - A\operatorname{dom}F)$ (since $\operatorname{dom}G = \mathcal{Y}$), so that by the Rockafellar–Fenchel duality theorem [4, Thm. 4.4.3] we have the strong duality

$$\sup_{\phi\in\mathcal{X}} -F(\phi) - G(A\phi) = \inf_{w\in\mathcal{Y}'} F^*(A^*w) + G^*(-w), \tag{9.47}$$

and a minimizer $w$ of the right-hand side exists unless the above equals $-\infty$ ($\mathcal{Y}' = \mathcal{M}(\overline{\Omega}\times[0,M])\times\mathcal{M}(\overline{\Omega}\times[0,M];\mathbf{R}^2\times\mathbf{R})$ here denotes the dual space to $\mathcal{Y}$, and $F^*$ and $G^*$ denote the convex conjugates of $F$ and $G$). As calculated before, the left-hand side equals $\sup_{\phi\in\tilde{\mathcal{K}}}\inf_{v\in\mathcal{C}}\int_{\overline{\Omega}\times[0,M]}\phi\cdot\mathrm{d}Dv$, so it remains to show that $\inf_{w\in\mathcal{Y}'}F^*(A^*w)+G^*(-w) = \inf_{v\in\mathcal{C}}\sup_{\phi\in\tilde{\mathcal{K}}}\int_{\overline{\Omega}\times[0,M]}\phi\cdot\mathrm{d}Dv$. We have

$$\begin{aligned} F^*(A^*w) &= \sup_{\phi\in\mathcal{X}}\langle\phi,A^*w\rangle - \iota_{\tilde{\mathcal{K}}}(\phi) = \sup_{\phi\in\mathcal{X}}\langle A\phi,w\rangle - \iota_{\tilde{\mathcal{K}}}(\phi)\\ &= \sup_{\phi\in\tilde{\mathcal{K}}\cap\mathcal{X}} \int_{\overline{\Omega}\times[0,M]} \operatorname{div}\phi\,\mathrm{d}w_1 + \int_{\overline{\Omega}\times[0,M]}\phi\cdot\mathrm{d}w_2, \end{aligned} \tag{9.48}$$

$$G^*(w) = \iota_{\mathcal{S}_1}(w_1) + \iota_{\mathcal{S}_2}(w_2) \tag{9.49}$$

with the sets

$$\mathcal{S}_1 = \{\mu\in\mathcal{M}(\overline{\Omega}\times[0,M]) \mid \mu\ll\mathcal{L}^3,\ 0\le\mu\le 1\}, \tag{9.50}$$

$$\mathcal{S}_2 = \{-1_{u(\mu_+,\mu_-)}n\,\mathcal{H}^2\llcorner\partial(\Omega\times(0,M))\}. \tag{9.51}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Thus $F^*(A^*w) + G^*(-w) \geq 0$ for all $w \in \mathcal{Y}'$, so that the infimum over all $w$ is finite, and $\inf_{w\in\mathcal{Y}'} F^*(A^*w) + G^*(-w) = \min_{w\in\mathcal{Y}'} F^*(A^*w) + G^*(-w)$. Furthermore, we obtain

$$\begin{aligned}
&\min_{w\in\mathcal{Y}'} F^*\big(A^*w\big) + G^*(-w) \\
&\quad = \min_{-w\in\mathcal{S}_1\times\mathcal{S}_2} \sup_{\phi\in\tilde{\mathcal{K}}\cap\mathcal{X}} \int_{\overline{\Omega}\times[0,M]} \operatorname{div}\phi \,\mathrm{d}w_1 + \int_{\overline{\Omega}\times[0,M]} \phi\cdot \mathrm{d}w_2 \\
&\quad = \min_{w_1\in L^1(\Omega\times(0,M);[0,1])} \sup_{\phi\in\tilde{\mathcal{K}}\cap\mathcal{X}} \int_{\partial(\Omega\times(0,M))} 1_{u(\mu_+,\mu_-)}\phi\cdot n\,\mathrm{d}\mathcal{H}^2 \\
&\qquad - \int_{\Omega\times(0,M)} w_1 \operatorname{div}\phi \,\mathrm{d}x\,\mathrm{d}s. \qquad (9.52)
\end{aligned}$$

Now the supremum on the right-hand side is only finite if $w_1$ is nonincreasing in the $s$-direction. Indeed, the finiteness of the supremum implies $\int_{\Omega\times(0,M)} w_1\partial_s\zeta \,\mathrm{d}x\,\mathrm{d}s \geq 0$ for all $\zeta \in C_0^\infty(\Omega\times(0,M);[0,\infty))$ since otherwise

$$\sup_{\phi\in\tilde{\mathcal{K}}\cap\mathcal{X}} -\int_{\Omega\times(0,M)} w_1 \operatorname{div}\phi \,\mathrm{d}x\,\mathrm{d}s \geq \sup_{\lambda>0} -\int_{\Omega\times(0,M)} w_1 \operatorname{div}(0,0,\lambda\zeta)\,\mathrm{d}x\,\mathrm{d}s = \infty.$$

The fundamental lemma of the calculus of variations now implies that $w_1$ is nonincreasing in the $s$-direction. Therefore by approximating $w_1$ with its mollifications it is straightforward to see that $\int_{\Omega\times(0,M)} w_1\partial_s\zeta \,\mathrm{d}x\,\mathrm{d}s \leq \mathcal{L}^2(\Omega)$ for any $\zeta \in C_0^\infty(\Omega\times(0,M);[-1,1])$. As a consequence, we have

$$\begin{aligned}
&\sup_{\phi\in\tilde{\mathcal{K}}\cap\mathcal{X}} \int_{\partial(\Omega\times(0,M))} 1_{u(\mu_+,\mu_-)}\phi\cdot n\,\mathrm{d}\mathcal{H}^2 - \int_{\Omega\times(0,M)} w_1 \operatorname{div}\phi\,\mathrm{d}x\,\mathrm{d}s \\
&\quad \geq \sup_{\phi\in\tilde{\mathcal{K}}\cap C_0^\infty(\Omega\times(0,M);\mathbf{R}^3)} -\int_{\Omega\times(0,M)} w_1 \operatorname{div}\phi\,\mathrm{d}x\,\mathrm{d}s \\
&\quad \geq \sup_{\phi\in\tilde{\mathcal{K}}\cap C_0^\infty(\Omega\times(0,M);\mathbf{R}^3),\zeta\in C_0^\infty(\Omega\times(0,M);[\frac{\tau(M)}{M},\frac{\tau(M)}{M}])} \\
&\qquad -\int_{\Omega\times(0,M)} w_1 \operatorname{div}\big(\phi+(0,0,\zeta)\big)\,\mathrm{d}x\,\mathrm{d}s - \frac{\tau(M)}{M}\mathcal{L}^2(\Omega) \\
&\quad \geq \sup_{\psi\in C_0^\infty(\Omega\times(0,M);\mathbf{R}^3),|\psi|\leq\frac{\tau(M)}{M}} -\int_{\Omega\times(0,M)} w_1 \operatorname{div}\psi\,\mathrm{d}x\,\mathrm{d}s - \frac{\tau(M)}{M}\mathcal{L}^2(\Omega) \\
&\quad = \frac{\tau(M)}{M}|w_1|_{\mathrm{TV}} - \frac{\tau(M)}{M}\mathcal{L}^2(\Omega), \qquad (9.53)
\end{aligned}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

where $|\cdot|_{\mathrm{TV}}$ denotes the total-variation seminorm. Thus the supremum is only finite if $w_1 \in \mathrm{BV}(\Omega \times (0,1))$, so that we may write

$$\begin{aligned}
&\min_{w \in \mathcal{Y}'} F^*(A^* w) + G^*(-w) \\
&\quad = \min_{w_1 \in \mathrm{BV}(\Omega\times(0,M);[0,1])} \sup_{\phi \in \tilde{\mathcal{K}} \cap \mathcal{X}} \int_{\partial(\Omega\times(0,M))} 1_{u(\mu_+,\mu_-)} \phi \cdot n \, \mathrm{d}\mathcal{H}^2 - \int_{\Omega\times(0,M)} w_1 \operatorname{div} \phi \, \mathrm{d}x \, \mathrm{d}s \\
&\quad = \min_{w_1 \in \mathrm{BV}(\Omega\times(0,M);[0,1])} \sup_{\phi \in \tilde{\mathcal{K}} \cap \mathcal{X}} \int_{\partial(\Omega\times(0,M))} (1_{u(\mu_+,\mu_-)} - w_1) \phi \cdot n \, \mathrm{d}\mathcal{H}^2 + \int_{\Omega\times(0,M)} \phi \cdot \mathrm{d}Dw_1 \\
&\quad = \min_{v \in \mathcal{C}} \sup_{\phi \in \tilde{\mathcal{K}} \cap \mathcal{X}} \int_{\overline{\Omega}\times[0,M]} \phi \cdot \mathrm{d}Dv,
\end{aligned} \tag{9.54}$$

where by density we may replace $\tilde{\mathcal{K}} \cap \mathcal{X}$ with $\tilde{\mathcal{K}}$. □

**Remark 9.2.2** (Predual variables of reduced regularity). Since the predual objective functional $\mathcal{D}$ and the functional $\phi \mapsto \int_{\overline{\Omega}\times[0,M]} \phi \cdot \mathrm{d}Dv$ for $v \in \mathcal{C}$ are continuous with respect to the norm $\|\phi\|_{L^1} + \|\operatorname{div}\phi\|_{L^1}$, throughout the statement of Theorem 9.2.3, all the suprema may actually be taken over

$$\begin{aligned}
\hat{\mathcal{K}} = \Bigg\{ &\phi = (\phi^x, \phi^s) \in L^1(\Omega \times (0,M); \mathbf{R}^2 \times [0,\infty)) \;\Big|\; \operatorname{div}\phi \in L^1(\Omega \times \mathbf{R}),\ \phi^s(x,s) \geq 0, \\
&|\phi^x(x,s)| \leq \tau'(0) \ \forall\, (x,s) \in \Omega \times (0,M), \\
&\left| \int_{s_1}^{s_2} \phi^x(x,s) \mathrm{d}s \right| \leq \tau(s_2 - s_1) \ \forall\, x \in \Omega, 0 \leq s_1 < s_2 \leq M \Bigg\}.
\end{aligned} \tag{9.55}$$

## 9.2.3 Calibrations for simple network configurations

Even without knowing equality in (9.8), we can make use of this inequality to prove the optimality of given transport networks by providing a so-called calibration (which is a predual certificate in the language of convex optimization). In fact, this was the aim of introducing the functional lifting of $J$ in [1] (the authors even considered a Mumford–Shah inpainting setting as we have it here). In this section, we provide calibrations for two exemplary transport networks, thereby showing the optimality of these network configurations and equality in (9.8) for these cases. Throughout we will use the notation of the previous section.

**Lemma 9.2.1** (Predual optima). *For any $\phi \in \hat{\mathcal{K}}$, there exists a divergence-free $\hat{\phi} \in \hat{\mathcal{K}}$ with no smaller predual cost $\mathcal{D}$. Thus, in the predual problem $\sup_{\phi \in \hat{\mathcal{K}}} \mathcal{D}(\phi)$, we may restrict to divergence-free vector fields $\phi$.*

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*Proof.* Let $\phi \in \hat{\mathcal{K}}$. By Smirnov's decomposition theorem [34, Thm. B–C] there exist a set $S$ of simple oriented curves of finite length (that is, measures of the form $\gamma_\# \dot\gamma \mathcal{H}^1 \llcorner [0,1]$ for $\gamma : [0,1] \to \overline{\Omega} \times [0,M]$ injective and Lipschitz, where $f_\# \mu$ denotes the pushforward of a measure $\mu$ under a map $f$) and a nonnegative measure $\rho$ on $S$ such that

$$\phi = \int_S \tilde\phi \,\mathrm{d}\rho(\tilde\phi), \quad \|\phi\|_{L^1} = \int_S \|\tilde\phi\|_{\mathcal{M}} \,\mathrm{d}\rho(\tilde\phi), \quad \|\operatorname{div}\phi\|_{\mathcal{M}} = \int_S \|\operatorname{div}\tilde\phi\|_{\mathcal{M}} \,\mathrm{d}\rho(\tilde\phi) \tag{9.56}$$

(the first equation means $\langle \phi, \psi\rangle = \int_S \langle \tilde\phi, \psi\rangle \,\mathrm{d}\rho(\tilde\phi)$ for every smooth test vector field $\psi$, and $\langle\cdot,\cdot\rangle$ is the duality pairing between Radon measures and continuous functions). Now consider

$$\tilde S = \{\gamma_\# \dot\gamma \mathcal{H}^1 \llcorner [0,1] \in S \mid \gamma(0) \in \Omega \times (0,M) \text{ or } \gamma(1) \in \Omega \times (0,M)\} \tag{9.57}$$

(which is $\rho$-measurable in the above sense). Then $\hat\phi = \int_{S\setminus\tilde S} \tilde\phi \,\mathrm{d}\rho(\tilde\phi)$ is divergence-free with $\mathcal{D}(\hat\phi) \le \mathcal{D}(\phi)$ and $\hat\phi \in \hat{\mathcal{K}}$. □

The previous lemma suggests to focus on divergence-free predual certificates, which in this context are called calibrations.

**Lemma 9.2.2.** *If there exists a divergence-free predual certificate for $v \in \mathcal{C}$, that is, a vector field $\hat\phi \in \hat{\mathcal{K}}$ with*

$$\operatorname{div}\hat\phi = 0 \quad \textit{and} \quad \tilde{\mathcal{G}}(v) = \int_{\overline{\Omega}\times[0,M]} \hat\phi \cdot \mathrm{d}Dv, \tag{9.58}$$

*then $v$ minimizes $\tilde{\mathcal{G}}$ on $\mathcal{C}$. Moreover, $\hat\phi$ is a predual certificate for any minimizer. In particular, if $v = 1_u$ for some $u \in \mathcal{A}_u$ and thus $\mathcal{E}(\mathcal{F}_u) = \tilde{\mathcal{E}}(u) = \mathcal{G}(1_u) = \tilde{\mathcal{G}}(1_u) = \int_{\overline{\Omega}\times[0,M]} \hat\phi \cdot \mathrm{d}D1_u$, then $u$ minimizes $\mathcal{E}(\mathcal{F}_u) = \tilde{\mathcal{E}}(u) = \mathcal{G}(1_u)$ over $\mathcal{A}_u$, and $\hat\phi$ is called a* calibration *for $u$.*

*Proof.* By weak duality from Theorem 9.2.3 we have $\tilde{\mathcal{G}}(v) \ge \mathcal{D}(\hat\phi)$ with equality if and only if $v$ and $\hat\phi$ are optimal. However, since $\hat\phi$ is divergence-free, we have

$$\mathcal{D}(\hat\phi) = \int_{\partial(\Omega\times(0,M))} 1_{u(\mu_+,\mu_-)} \phi \cdot n \,\mathrm{d}\mathcal{H}^2 - \int_{\Omega\times(0,M)} v \operatorname{div}\phi \,\mathrm{d}x\,\mathrm{d}s = \int_{\overline{\Omega}\times[0,M]} \hat\phi \cdot \mathrm{d}Dv = \tilde{\mathcal{G}}(v) \tag{9.59}$$

after an integration by parts; thus $v \in \mathcal{C}$ is minimizing, and $\hat\phi \in \hat{\mathcal{K}}$ is maximizing. Now any other minimizer $\tilde v \in \mathcal{C}$ satisfies $\tilde{\mathcal{G}}(\tilde v) = \tilde{\mathcal{G}}(v) = \mathcal{D}(\hat\phi) = \int_{\overline{\Omega}\times[0,M]} \hat\phi \cdot \mathrm{d}D\tilde v$ by the same calculation, so that $\hat\phi$ also calibrates $\tilde v$. □

**Remark 9.2.3** (Sequences as calibrations and less regularity). By an obvious modification of the above argument, the existence of the divergence-free $\hat\phi \in \hat{\mathcal{K}}$ can be replaced by the existence of a sequence $\phi_1, \phi_2, \ldots \in \hat{\mathcal{K}}$ of divergence-free vector fields such with $\tilde{\mathcal{G}}(v) = \lim_{n\to\infty} \int_{\overline{\Omega}\times[0,M]} \phi_n \cdot \mathrm{d}Dv$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

In the remainder of the section, we provide two examples for calibrations, one for a classic network configuration that can be and has been analyzed classically on the level of graphs and one that cannot be analyzed on such a basis. We begin by proving the angle conditions for triple junctions, which, as mentioned above, can also easily be obtained by a vertex perturbation argument. Any triple junction can locally be interpreted as having a single source point and two sink points (or vice versa), which we do below.

**Example 9.2.2** (Triple junction). Let a point source and two point sinks be located on the boundary of the unit disk $\Omega$,

$$\mu_+ = (m_1 + m_2)\delta_{-e_0}, \quad \mu_- = m_1\delta_{e_1} + m_2\delta_{e_2} \quad \text{for } e_0, e_1, e_2 \in \partial\Omega = S^1,\ m_1, m_2 > 0, \tag{9.60}$$

where the vectors $e_0$, $e_1$, $e_2$ satisfy the angle condition

$$0 = \tau(m_1)e_1 + \tau(m_2)e_2 - \tau(m_1 + m_2)e_0. \tag{9.61}$$

Then the mass flux $\mathcal{F} = (m_1 + m_2)e_0\mathcal{H}^1 \llcorner e_0 + m_1e_1\mathcal{H}^1 \llcorner e_1 + m_2e_2\mathcal{H}^1 \llcorner e_2$ minimizes $\mathcal{E}$ on $\mathcal{A}_{\mathcal{F}}$.

To prove this statement, assume without loss of generality that $e_i = (\cos\varphi_i, \sin\varphi_i)$, $i = 0, 1, 2$, with $0 \le \varphi_1 \le \varphi_0 \le \varphi_2 \le 2\pi$ (see Figure 9.3, left). Then $u_{\mathcal{F}}$ reads

$$u_{\mathcal{F}}(r, \varphi) = \begin{cases} 0 & \text{if } \varphi \in [\varphi_2, 2\pi - \varphi_0), \\ m_2 & \text{if } \varphi \in [\varphi_1, \varphi_2), \\ m_1 + m_2 & \text{otherwise} \end{cases} \tag{9.62}$$

in polar coordinates, and its maximum is $M = m_1 + m_2$. Now set

$$\phi(x, s) = \begin{cases} -\frac{\tau(m_2)}{m_2}(e_2^\perp, 0) & \text{if } 0 \le s \le m_2, \\ -\frac{\tau(m_1)}{m_1}(e_1^\perp, 0) & \text{if } m_2 \le s \le M, \end{cases} \tag{9.63}$$

where $\perp$ denotes the counterclockwise rotation by $\pi/2$. With this choice, we have

$$\begin{aligned} \int_{\overline{\Omega}\times[0,M]} \phi \cdot \mathrm{d}D1_{u_{\mathcal{F}}} &= \tau(m_1)(1 + e_0^\perp \cdot e_1^\perp) + \tau(m_2)(1 + e_0^\perp \cdot e_2^\perp) \\ &= \tau(m_1) + \tau(m_2) + \tau(m_1 + m_2) = \mathcal{E}(\mathcal{F}), \end{aligned} \tag{9.64}$$

where in the second equality, we used the (inner product with $e_0$ of the) angle condition. Furthermore, $\operatorname{div}\phi = 0$, and we have $|\int_{s_1}^{s_2} \phi^x(x, s)\,\mathrm{d}s| \le \tau(s_2 - s_1)$ for all $0 \le s_1 \le s_2 \le M$, Indeed, for $s_2 \le m_2$ or $s_1 \ge m_2$, this is trivial to check, and for $s_1 \le m_2 \le s_2$, we

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

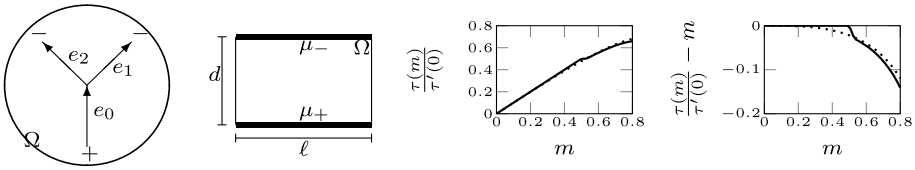



**Figure 9.3:** Left: Illustration of the notation in Example 9.2.2. Right: Illustration of the notation in Example 9.2.3 and the lower bound on $\frac{\tau(m)}{\tau'(0)}$ for $\beta = 1$ (solid) and $\beta = 1.05$ (dotted). Any transportation cost $\tau$ yields a diffuse flux if $\frac{\tau(m)}{\tau'(0)}$ lies above the solid or the dotted curve (note that it automatically lies below the diagonal). For better visualization, we also show the lower bound on $\frac{\tau(m)}{\tau'(0)} - m$.

set $\alpha = \min\{\frac{m_2 - s_1}{m_2}, \frac{s_2 - m_2}{m_1}\}$ and calculate

$$
\begin{aligned}
&\left|\int_{s_1}^{s_2} \phi^x(x,s)\,\mathrm{d}s\right| \\
&\quad= \left|(m_2 - s_1)\frac{\tau(m_2)}{m_2}e_2 + (s_2 - m_2)\frac{\tau(m_1)}{m_1}e_1\right| \\
&\quad= \left|\alpha\big(\tau(m_2)e_2 + \tau(m_1)e_1\big) + \left(\frac{m_2 - s_1}{m_2} - \alpha\right)\tau(m_2)e_2 + \left(\frac{s_2 - m_2}{m_1} - \alpha\right)\tau(m_1)e_1\right| \\
&\quad\le \alpha\tau(m_1 + m_2) + \left(\frac{m_2 - s_1}{m_2} - \alpha\right)\tau(m_2) + \left(\frac{s_2 - m_2}{m_1} - \alpha\right)\tau(m_1) \\
&\qquad + \left(1 + \alpha - \frac{m_2 - s_1}{m_2} - \frac{s_2 - m_2}{m_1}\right)\tau(0) \\
&\quad\le \tau\bigg(\alpha(m_1 + m_2) + \left(\frac{m_2 - s_1}{m_2} - \alpha\right)m_2 + \left(\frac{s_2 - m_2}{m_1} - \alpha\right)m_1 \\
&\qquad + \left(1 + \alpha - \frac{m_2 - s_1}{m_2} - \frac{s_2 - m_2}{m_1}\right)0\bigg) \\
&\quad= \tau(s_2 - s_1),
\end{aligned}
\tag{9.65}
$$

where in the first inequality, we used the triangle inequality and the angle condition, and in the last inequality, we used Jensen's inequality with convex combination coefficients $\alpha$, $(\frac{m_2 - s_1}{m_2} - \alpha)$, $(\frac{s_2 - m_2}{m_1} - \alpha)$, $(1 + \alpha - \frac{m_2 - s_1}{m_2} - \frac{s_2 - m_2}{m_1})$. Thus $\phi \in \hat{\mathcal{K}}$, as desired.

The second example shows that even for strictly concave transportation cost $\tau$, we may have a diffuse flux without network formation.

**Example 9.2.3** (Diffuse flux). Let the source and sink be two line measures opposite of each other, that is,

$$
\begin{aligned}
&\mu_+ = m\mathcal{H}^1 \llcorner [0,\ell] \times \{0\}, \\
&\mu_- = m\mathcal{H}^1 \llcorner [0,\ell] \times \{d\} \quad \text{for some } m, d, \ell > 0 \text{ and } \Omega = (0,\ell) \times (0,d)
\end{aligned}
\tag{9.66}
$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

(see Figure 9.3, right). By rescaling space and the transportation cost we may reduce the setting to the equivalent one with $m = d = 1$ without loss of generality. If the transportation cost $\tau$ satisfies

$$\frac{\tau(m)}{\tau'(0)} \geq \max\left\{\min\left\{\frac{m}{\beta}, \frac{1}{2}\right\}, \sqrt{\beta^2 - m^2}\,\operatorname{arsinh}\frac{m}{\sqrt{\beta^2 - m^2}}\right\} \quad \forall m < \beta \tag{9.67}$$

for some $\beta \geq 1$ (note that necessarily $\tau(m)/\tau'(0) \leq m$), then the optimal flux is given by the diffuse $\mathcal{F} = (0,1)\mathcal{L}^2 \llcorner \Omega$.

Note that for $\beta$ large enough, the above bound on $\tau$ simply evaluates to the strictly concave $\sqrt{\beta^2 - m^2}\,\operatorname{arsinh}\frac{m}{\sqrt{\beta^2-m^2}}$; from then on, any larger $\beta$ produces a weaker bound.

To prove the statement, note that $u_{\mathcal{F}}(x) = x_1$ with maximum $M = \ell$ and set

$$\begin{aligned}
\psi(x) &= \begin{cases} \frac{-x^\perp}{|x|} & \text{if } |x| \leq \frac{1}{2}, \\ 0 & \text{else,} \end{cases} \qquad \tilde{\psi}(x) = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\beta} \end{pmatrix} \psi\left(\begin{pmatrix} \frac{1}{\beta} & 0 \\ 0 & 1 \end{pmatrix} x\right), \\
\phi(x,s) &= \begin{cases} (\tilde{\psi}(x_1 - s, x_2), 0) & \text{if } x_2 \leq \frac{1}{2}, \\ (\tilde{\psi}(s - x_1, 1 - x_2), 0) & \text{else} \end{cases}
\end{aligned} \tag{9.68}$$

(note that for each $s$, $\phi$ is symmetric about $x_2 = \frac{1}{2}$, describing an elliptic flow in each half). It is straightforward to check that

$$\mathcal{E}(\mathcal{F}) = \ell = \int_{\overline{\Omega}\times[0,M]} \phi \cdot \mathrm{d}D1_{u_{\mathcal{F}}} \tag{9.69}$$

and $\operatorname{div}\phi = 0$. Furthermore, we need to check the condition $|\int_{s_1}^{s_2} \phi^x(0, x_2, s)\,\mathrm{d}s| \leq \tau(s_2 - s_1)$ for all $-\beta\sqrt{1/4 - x_2^2} \leq s_1 \leq s_2 \leq \beta\sqrt{1/4 - x_2^2}$ (outside this range, $\phi$ is zero anyway), where due to symmetry it suffices to consider the position $x_1 = 0$. We can calculate (without loss of generality, for $x_2 \leq \frac{1}{2}$)

$$\begin{aligned}
\left|\int_{s_1}^{s_2} \phi^x(0, x_2, s)\,\mathrm{d}s\right| &= \left|\begin{pmatrix} 1 & 0 \\ 0 & 1/\beta \end{pmatrix} \int_{s_1}^{s_2} \psi(-s/\beta, x_2)\,\mathrm{d}s\right| = \left|\int_{s_1}^{s_2} \frac{\binom{x_2}{s/\beta^2}}{\sqrt{x_2^2 + s^2/\beta^2}}\,\mathrm{d}s\right| \\
&= \left|x_2 \begin{pmatrix} \beta(\operatorname{arsinh} z_2 - \operatorname{arsinh} z_1) \\ \sqrt{1 + z_2^2} - \sqrt{1 + z_1^2} \end{pmatrix}\right| \\
&= x_2 \sqrt{\beta^2(\operatorname{arsinh} z_2 - \operatorname{arsinh} z_1)^2 + \left(\sqrt{1 + z_2^2} - \sqrt{1 + z_1^2}\right)^2}
\end{aligned} \tag{9.70}$$

for $z_i = s_i/(x_2\beta)$, $i = 1, 2$. Let us abbreviate this function by $f(s_1, s_2, x_2, \beta)$. We need to have $\tau(m) \geq f(s - \frac{m}{2}, s + \frac{m}{2}, x_2, \beta)$ for any choice of $s$ (which due to symmetry we may assume to be nonnegative) and $x_2$. Now it turns out that $f(s - \frac{m}{2}, s + \frac{m}{2}, x_2, \beta)$ has no

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

critical points as a function of $s$ and $x_2$. Indeed,

$$\frac{\mathrm{d}}{\mathrm{d}x_2} f\Big(s-\frac{m}{2}, s+\frac{m}{2}, x_2, \beta\Big) = -x_2\beta^2 \frac{\begin{array}{c}(\operatorname{arsinh} z_2 - \operatorname{arsinh} z_1)^2 - (\operatorname{arsinh} z_2 - \operatorname{arsinh} z_1)\big(\frac{z_2}{\sqrt{1+z_2^2}} - \frac{z_1}{\sqrt{1+z_1^2}}\big) \\ + \big(2 - \sqrt{\frac{1+z_2^2}{1+z_1^2}} - \sqrt{\frac{1+z_1^2}{1+z_2^2}}\big)/\beta^2\end{array}}{f(s-\frac{m}{2}, s+\frac{m}{2}, x_2, \beta)}, \tag{9.71}$$

$$\frac{\mathrm{d}}{\mathrm{d}s} f\Big(s-\frac{m}{2}, s+\frac{m}{2}, x_2, \beta\Big) = \beta \frac{(\operatorname{arsinh} z_2 - \operatorname{arsinh} z_1)\big(\frac{1}{\sqrt{1+z_2^2}} - \frac{1}{\sqrt{1+z_1^2}}\big) + \big(\sqrt{1+z_2^2} - \sqrt{1+z_1^2}\big)\big(\frac{z_2}{\sqrt{1+z_2^2}} - \frac{z_1}{\sqrt{1+z_1^2}}\big)/\beta^2}{f(s-\frac{m}{2}, s+\frac{m}{2}, x_2, \beta)}, \tag{9.72}$$

and we can check that there are no joint zeros $(z_1, z_2)$ of both expressions. Consequently, $f(s-\frac{m}{2}, s+\frac{m}{2}, x_2, \beta)$ becomes extremal on the boundary of the admissible domain $\{(s, x_2) \in \mathbf{R} \times [0,\infty) \mid \frac{(s\pm m/2)^2}{\beta^2} + x_2^2 \le \frac{1}{4}\}$ (such that $s_2 = s + \frac{m}{2} \le \beta\sqrt{1/4 - x_2^2}$). We can readily evaluate $f(s-\frac{m}{2}, s+\frac{m}{2}, 0, \beta) = \frac{||s_2|-|s_1||}{\beta} \le \min\{\frac{m}{\beta}, \frac{1}{2}\}$, and thus we require $\tau(m) \ge \min\{\frac{m}{\beta}, \frac{1}{2}\}$. On the other boundary, $x_2 = \sqrt{\frac{1}{4} - \frac{(s\pm m/2)^2}{\beta^2}}$, for symmetry reasons, it suffices to consider $s \ge 0$. It turns out that $f(s-\frac{m}{2}, s+\frac{m}{2}, \sqrt{\frac{1}{4} - \frac{(s+m/2)^2}{\beta^2}}, \beta)$ is initially decreasing in $s \in [0, \frac{\beta-m}{2}]$ and may then again increase, depending on the size of $\beta$ and $m$. Thus the maximum value is taken either at $s = \frac{\beta-m}{2}$ (which is the case $x_2 = 0$ already treated above) or at $s = 0$. Hence we additionally need $\tau(m) \ge f(-\frac{m}{2}, \frac{m}{2}, \sqrt{\frac{1}{4} - \frac{m^2}{4\beta^2}}, \beta) = \sqrt{\beta^2 - m^2} \operatorname{arsinh} \frac{m}{\sqrt{\beta^2 - m^2}}$ so that $\phi \in \hat{\mathcal{K}}$ as desired.

## 9.3 Adaptive finite elements for functional lifting problems

Convex optimization problems arising from functional lifting as introduced in Section 9.2.2 require a careful numerical treatment due to several reasons. First, the lifted problem has an objective variable living in three rather than two space dimensions, which requires a careful discretization to provide a straightforward translation between the two- and three-dimensional models. Furthermore, the problem size is strongly increased by the lifting; not only do the variables live in a higher-dimensional space, but also the set $\hat{\mathcal{K}}$ has a constraint for every $(x, s_1, s_2) \in \Omega \subset \mathbf{R}^2 \times \mathbf{R} \times \mathbf{R}$, so that

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

the problem essentially behaves like a four-dimensional one. Finally, to make the algorithm reliable and avoid unwanted effects, the discretization of the feasible set $\hat{\mathcal{K}}$ should be feasible itself (that is, a subset of $\hat{\mathcal{K}}$), which means that we must be able to reduce the infinite number of nonlocal constraints in $\hat{\mathcal{K}}$ to a finite number.

One possible way to jointly tackle the previously mentioned challenges is a adaptive finite element approach defined on grids consisting of prism-shaped elements. As before, to emphasize the difference between the original image domain and range, for a point in $\Omega \times (0, M)$, we denote its first two coordinates as $x$-coordinates and the third one as an $s$-coordinate with respect to the standard basis of $\mathbf{R}^3$.

## 9.3.1 Adaptive triangular prism grids

We start by recalling the definition of a two-dimensional simplicial grid (see, for instance, [36]).

**Definition 9.3.1** (Simplex, simplicial grid in 2D). A *two-dimensional simplex* $(x^0, x^1, x^2)$ is a 3-tuple with nodes $x^0, x^1, x^2 \in \mathbf{R}^2$ that do not lie on a one-dimensional hyperplane. The convex hull $\operatorname{conv}\{x^0, x^1, x^2\}$ is also denoted as a simplex. A *two-dimensional simplicial grid* on $\overline{\Omega}$ is a set of two-dimensional simplices with pairwise disjoint interior and union $\overline{\Omega}$.

Based on a two-dimensional simplicial grid for the image domain $\Omega$, we define a *lifted counterpart* consisting of triangular prism-shaped elements. For two tuples $(x^0, \dots, x^k)$ and $(s^0, \dots, s^l)$ of points in $\mathbf{R}^n$ and $\mathbf{R}^m$, we will write $(x^0, \dots, x^k) \times (s^0, \dots, s^l)$ for the tuple $((x^0, s^0), (x^1, s^0), \dots (x^k, s^l))$ of points in $\mathbf{R}^n \times \mathbf{R}^m$.

**Definition 9.3.2** (Triangular prism element). A *triangular prism element* $T$ is a 6-tuple $T_x \times T_s$ of nodes in $\mathbf{R}^3$, where $T_x = (x^0, x^1, x^2)$ is a two-dimensional simplex, and $T_s = (s^0, s^1)$ for $s^0, s^1 \in \mathbf{R}$ with $s^1 > s^0$. If there is no ambiguity, then the convex hull of the nodes is also denoted a triangular prism element (and $T_x$ and $T_s$ are likewise identified with their corresponding convex hulls). The *vertical* and *horizontal edges* of $T$ are given by $x^i \times (s^0, s^1)$ and $(x^i, x^j) \times s^k$, respectively, for $i, j \in \{0, 1, 2\}$, $k \in \{0, 1\}$. We similarly define its *vertical* and *horizontal faces*.

A single triangular prism element can be refined either in the $x$-plane or in the $s$-direction as illustrated in Figure 9.4, where we suggest to use the obvious extension of the standard bisection method for a two-dimensional simplicial grid (see, for instance, [36]).

**Definition 9.3.3** (Element refinement). The *s-refinement* of a triangular prism element $T = (x^0, x^1, x^2) \times (s^0, s^1)$ is the pair of triangular prism elements

$$(x^0, x^1, x^2) \times \left(s^0, \frac{s^0 + s^1}{2}\right) \quad \text{and} \quad (x^0, x^1, x^2) \times \left(\frac{s^0 + s^1}{2}, s^1\right). \tag{9.73}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use



**Figure 9.4:** Subdivision of a triangular prism element $T$ by $x$-refinement along the longest (bold) horizontal edge (left) and $s$-refinement(right).

Assuming without loss of generality $(x^1, x^2)$ to be the longest edge of $(x^0, x^1, x^2)$, the *x-refinement* of $T$ is the pair of triangular prism elements

$$\left(x^0, x^1, \frac{x^1 + x^2}{2}\right) \times (s^0, s^1) \quad \text{and} \quad \left(x^0, \frac{x^1 + x^2}{2}, x^2\right) \times (s^0, s^1). \tag{9.74}$$

We aim for simulations on an adaptively refined grid. During refinement, we want to keep a certain regularity condition of the grid, which we call semiregular.

**Definition 9.3.4** (Triangular prism grid and hanging nodes). A *triangular prism grid* $\mathcal{T}$ on $\overline{\Omega} \times [0, M]$ is a set of triangular prism elements with pairwise disjoint interior and union $\overline{\Omega} \times [0, M]$. Its *set of nodes* $\mathcal{N}(\mathcal{T})$ is the union of the nodes of all its elements.

A node $N \in \mathcal{N}(\mathcal{T})$ is called *hanging* if there is an element $T \in \mathcal{T}$ with $N \in T$, but $N$ is not a node of $T$. It is *s-hanging* (or *x-hanging*) if for any such $T$, the node $N$ lies on a vertical (or horizontal) edge of $T$.

The grid $\mathcal{T}$ is called *regular* if it does not contain any hanging nodes. It is called *semiregular* if it does not contain any $x$-hanging nodes and if any two elements $T = (x^0, x^1, x^2) \times (s^0, s^1)$, $S = (y^0, y^1, y^2) \times (r^0, r^1) \in \mathcal{T}$ with nonempty intersection either exactly share a node, an edge, or a face or satisfy either $s^0, s^1 \in \{r^0, \frac{r^0 + r^1}{2}, r^1\}$ or $r^0, r^1 \in \{s^0, \frac{s^0 + s^1}{2}, s^1\}$.

Obviously, in addition to sharing a full edge or face, neighboring elements in a semiregular prism grid may also be such that a vertical edge or face of one may be a vertical half-edge or half-face of the other, as illustrated in Figure 9.5, resulting in $s$-hanging nodes. The limitation of $s$-hanging nodes to one per edge is a natural

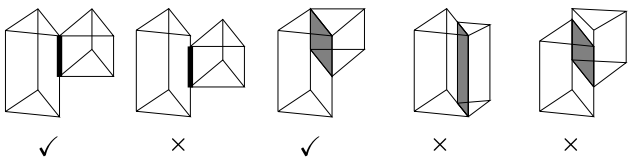



**Figure 9.5:** Examples of allowed and forbidden neighboring relations in a semiregular triangular prism grid.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

convention to prevent too many successive hanging nodes, which are typically not associated with any degrees of freedom. The $x$-refinement only allows bisection of the longest edge, which is the standard means to prevent degeneration of the interior element angles.

The rationale behind concentrating on semiregular grids is that these allow a simple discretization of the set of lifting constraints (as will be detailed in Section 9.3.2) and at the same time are sufficiently compatible with local refinement. Indeed, had we only admitted regular grids, then any $s$-refinement would have to be done globally for all elements in a two-dimensional cross-section of the grid, whereas the possibility of $s$-hanging nodes in semiregular grids allows us to subdivide just a few local elements in the $s$-direction. On the other hand, $x$-refinement can be done locally at a position $\hat{x}$ in the $x$-plane but has to be performed simultaneously for all elements along the $s$-coordinate sitting above $\hat{x}$. However, this is just global refinement along a one-dimensional direction (rather than the above-mentioned global refinement in a two-dimensional cross-section), and due to the possibility of local $s$-refinement, we practically only have quite few elements along this direction.

A suitable algorithm for grid refinement should preserve the semiregularity of the grid. Thus the refinement of one element potentially implies the successive refinement of several neighboring elements. In case of $x$-refinement, this affects all elements sharing a bisected face or edge with the refined element (that is, the element above and below as well as the neighbor across the subdivided vertical face). In case of $s$-refinement, the *half-edge* rule has to be maintained, such that horizontal neighbors whose heights exceed twice the height of the refined element need to be refined successively. It is a standard fact that the resulting chains of successive element refinements terminate after a finite number of steps.

Finally, we note that the projection of a semiregular triangular prism grid onto the $x$-hyperplane $\mathbf{R}^2 \times \{0\}$ naturally yields a two-dimensional simplicial grid by construction, and so does every horizontal slice of the grid.

## 9.3.2 Reduction of the constraint set $\hat{\mathcal{K}}$

Having fixed the grid, we now need to discretize functions on that grid. We will choose these functions to be piecewise linear in the $x$-direction and piecewise constant in the $s$-direction (the details are given in Section 9.3.3). In this section, we give the reason for that choice: It easily allows us to check and project onto the conditions in the convex set $\hat{\mathcal{K}}$. A priori, this is very challenging, since for every base point $x \in \Omega$, we have an infinite number of inequality constraints. Furthermore, after discretization, the inequality constraints for different base points might interdepend on each other in a nontrivial way due to interpolation between different nodal values. We first show that for functions piecewise constant along the lifting dimension, the infinite number of inequality constraints at each base point $x \in \Omega$ reduces to a finite number. We then

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

prove that if the functions are piecewise linear in the $x$-direction, then only the constraints for nodal base points have to be checked.

**Theorem 9.3.1** (Constraint set for functions piecewise constant in $s$). *Let $0 = t_0 < t_1 < \cdots < t_p = M$ be a partition of $[0, M]$, and let $\psi : [0, M) \to \mathbf{R}^2$ be piecewise constant,*

$$\psi(s) = C_i \quad \text{for } s \in [t_i, t_{i+1}),\ i = 0, \dots, p-1. \tag{9.75}$$

*Let $\tau$ be a transportation cost. Then we have*

$$\begin{aligned} &\left| \int_{s_1}^{s_2} \psi \, \mathrm{d}s \right| \le \tau(|s_2 - s_1|) \quad \forall s_1, s_2 \in [0, M] \quad \text{if and only if} \\ &\left| \int_{s_1}^{s_2} \psi \, \mathrm{d}s \right| \le \tau(|s_2 - s_1|) \quad \forall s_1, s_2 \in \{t_0, \dots, t_p\}. \end{aligned} \tag{9.76}$$

*Proof.* We only need to prove one implication (the other being trivial). Let $|\int_{s_1}^{s_2} \psi \, \mathrm{d}s| \le \tau(|s_2 - s_1|)$ for all $s_1, s_2 \in \{t_0, \dots, t_p\}$. Now fix arbitrary $s_1, s_2 \in [0, M]$, where without loss of generality we have $s_1 < s_2$. If $s_1, s_2 \in [t_i, t_{i+1}]$ for some $i \in \{0, \dots, p\}$, then

$$\left| \int_{s_1}^{s_2} \psi \, \mathrm{d}s \right| = (s_2 - s_1)|C_i| = \frac{s_2 - s_1}{t_{i+1} - t_i} \left| \int_{t_i}^{t_{i+1}} \psi \, \mathrm{d}s \right| \le \frac{s_2 - s_1}{t_{i+1} - t_i} \tau(t_{i+1} - t_i) \le \tau(s_2 - s_1) \tag{9.77}$$

due to $\tau(0) = 0$ and the concavity of $\tau$. It remains to consider the case $s_1 \in [t_i, t_{i+1}]$ and $s_2 \in [t_j, t_{j+1}]$ with $i < j$. To this end, consider the function $f : [t_i, t_{i+1}] \times [t_j, t_{j+1}] \to \mathbf{R}$,

$$f(s_1, s_2) = \left| \int_{s_1}^{s_2} \psi \, \mathrm{d}s \right| - \tau(s_2 - s_1) = \left| (t_{i+1} - s_1) C_i + \int_{t_{i+1}}^{t_j} \psi \, \mathrm{d}s + (s_2 - t_j) C_j \right| - \tau(s_2 - s_1). \tag{9.78}$$

As a composition of a convex with an affine function, $f$ is jointly convex in both arguments. Therefore, since $f \le 0$ at the four corners (the convex extreme points) of its domain, we have $f \le 0$ all over the domain, which finishes the proof. □

As a consequence, a piecewise constant approximation of the variables in the lifted direction allows an efficient constraint handling. This feature breaks down already for piecewise linear instead of piecewise constant functions (where it becomes much harder to check the constraints), as the following simple counterexample illustrates.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Example 9.3.1** (Constraint set for functions piecewise linear in $s$). Let $p \in \mathbb{N}$, $h_s = \frac{M}{p}$, and $t_i = ih_s$ for $i = 0, \dots, p$. Fix an arbitrary $C \in \mathbf{R}^2$ and define $\psi : [0, M] \to \mathbf{R}^2$ as

$$\psi(x) = \begin{cases} \frac{2C}{h_s}(s - t_i) - C & \text{if } s \in [t_i, t_{i+1}],\ i \text{ even}, \\ \frac{2C}{h_s}(t_i - s) + C & \text{if } s \in [t_i, t_{i+1}],\ i \text{ odd} \end{cases} \tag{9.79}$$

(see Figure 9.6). Then, obviously, $\left| \int_{t_i}^{t_j} \psi \, \mathrm{d}s \right| = 0$ for any $i, j \in \{0, \dots, p\}$, whereas for $\tilde{s} = \frac{t_i + t_{i+1}}{2}$, we have

$$\left| \int_{t_i}^{\tilde{s}} \psi \, \mathrm{d}s \right| = \frac{h_s}{4} |C|, \tag{9.80}$$

which can be arbitrarily large depending on $C$.

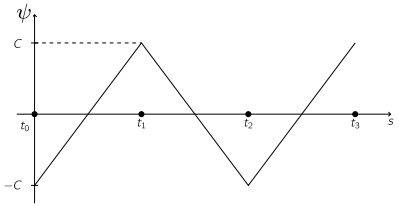



**Figure 9.6:** Sketch of the function $\psi$ from Example 9.3.1.

We next state that for piecewise linear discretization in the $x$-direction, it suffices to consider a finite number of base points.

**Theorem 9.3.2** (Constraint set for functions piecewise linear in $x$). *Let $\tilde{\mathcal{T}}$ be a regular two-dimensional simplex grid on $\overline{\Omega}$ with node set $\mathcal{N}(\tilde{\mathcal{T}})$, and let $\phi : \overline{\Omega} \times [0, M] \to \mathbf{R}^3$ be piecewise linear in the x-direction, that is, for each $s \in [0, M]$, the function $x \mapsto \phi(x, s)$ is continuous and affine on each simplex $T \in \tilde{\mathcal{T}}$. Then*

$$\begin{aligned} &\left| \int_{s_1}^{s_2} \phi^x(x, s) \, \mathrm{d}s \right| \le \tau(|s_2 - s_1|), \quad |\phi^x(x, s_1)| \le \tau'(0), \quad \textit{and} \\ &\qquad \phi^s(x, s_1) \ge 0 \quad \forall s_1, s_2 \in [0, M] \end{aligned} \tag{9.81}$$

*is satisfied for all $x \in \overline{\Omega}$ if and only if it is satisfied for all $x \in \mathcal{N}(\tilde{\mathcal{T}})$.*

*Proof.* Again, one implication is trivial, and we show the other one. Let the constraints be satisfied for all $x \in \mathcal{N}(\tilde{\mathcal{T}})$. Now pick an arbitrary $x \in \overline{\Omega}$ and let $T = (x^0, x^1, x^2) \in \tilde{\mathcal{T}}$ be such that $x \in T$ and thus $x = \lambda_0 x_0 + \lambda_1 x_1 + \lambda_2 x_2$ for convex combination coefficients

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$\lambda_0, \lambda_1, \lambda_2 \in [0,1]$. Now the function $\phi(x,\cdot)$ can be written as the convex combination $\phi(x,\cdot) = \lambda_0\phi(x_0,\cdot) + \lambda_1\phi(x_1,\cdot) + \lambda_2\phi(x_2,\cdot)$. Since the constraints are convex in $\phi(x,\cdot)$ and are satisfied for $\phi(x_0,\cdot)$, $\phi(x_1,\cdot)$, and $\phi(x_2,\cdot)$, they are also satisfied for $\phi(x,\cdot)$. □

Note that the important feature of the piecewise linear discretization in the $x$-direction, which allows the above constraint reduction, is that the nodal basis of each element is a nonnegative partition of unity and can thus at each point $x \in \overline{\Omega}$ be viewed as a set of convex combination coefficients. This feature breaks down for higher-order elements.

Summarizing, if the ($x$-component of the) flux $\phi$ is discretized as piecewise constant in the $s$-direction and piecewise linear in the $x$-direction, then the constraints forming the set $\hat{\mathcal{K}}$ only need to be checked at all nodes of the underlying grid.

The above also explains why we aim for semiregular grids and avoid $x$-hanging nodes: Otherwise, we would have to test the constraints also for all base points $\hat{x}$ that correspond to $x$-hanging nodes (and over these points, we would need to consider all $s_1, s_2 \in [0,M]$ at which there is an element face, not only those $s_1, s_2 \in [0,M]$ for which $(\hat{x}, s_1)$ and $(\hat{x}, s_2)$ are nodes). Furthermore, the projection of a discretized vector field $\phi$ onto the constraint set will be much more complicated: Without $x$-hanging nodes, we can perform the projection independently for all nodes of the underlying two-dimensional simplex grid. With $x$-hanging nodes, however, the constraints are no longer independent, since the function value at a hanging node is slaved to the function values at the neighboring nonhanging nodes.

### 9.3.3 Finite element discretization

We now aim to discretize our convex saddle point problem

$$\inf_{v\in\mathcal{C}} \sup_{\phi\in\hat{\mathcal{K}}} \int_{\overline{\Omega}\times[0,M]} \phi \cdot \mathrm{d}Dv \tag{9.82}$$

based on a triangular prism finite element approach. Motivated by Theorems 9.3.1 and 9.3.2, on a semiregular triangular prism grid $\mathcal{T}$, we define the discrete function spaces

$$S^1(\mathcal{T}) = \{w \in C^0(\overline{\Omega}\times[0,M]) \mid w|_T \text{ is affine for all } T \in \mathcal{T}\}, \tag{9.83}$$

$$\begin{aligned} S^{0,1}(\mathcal{T}) = \{w : \overline{\Omega}\times[0,M) \to \mathbf{R} \mid\ & w(\cdot,s) \in C^0(\overline{\Omega}) \text{ for all } s \in [0,M) \\ & \text{and for all } T \in \mathcal{T}, \text{ there are } a,b,c \in \mathbf{R} \\ & \text{with } w|_{T'}(x,s) = ax_1 + bx_2 + c\}, \end{aligned} \tag{9.84}$$

where $w|_T$ denotes the restriction of $w$ onto $T$, and where $T'$ denotes the triangular prism element without its upper triangular face. Obviously, on $T = (x^0,x^1,x^2)\times(s^0,s^1)$, any function $w \in S^1(\mathcal{T})$ is uniquely determined by its values at the element nodes,

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

whereas $w|_{T'}$ for $w \in S^{0,1}(\mathcal{T})$ is uniquely determined by the values of $w$ at the bottom nodes $(x^0, s^0)$, $(x^1, s^0)$, $(x^2, s^0)$. Consequently, any $w \in S^1(\mathcal{T})$ is uniquely determined by its function values at the set of all except the hanging nodes, which we denote by $\mathcal{N}'(\mathcal{T}) \subset \mathcal{N}(\mathcal{T})$, and any $w \in S^{0,1}(\mathcal{T})$ is uniquely determined by its function values at the set of all except the hanging and the top-most nodes, which we denote by $\mathcal{N}''(\mathcal{T}) \subset \mathcal{N}'(\mathcal{T})$. Numbering the nodes in $\mathcal{N}'(\mathcal{T})$ and $\mathcal{N}''(\mathcal{T})$ as $N_1, \dots, N_{q'}$ and $N_1, \dots, N_{q''}$, respectively, we can thus define a nodal basis $(\theta_1, \dots, \theta_{q'})$ of $S^1(\mathcal{T})$ and $(\psi_1, \dots, \psi_{q''})$ of $S^{0,1}(\mathcal{T})$ via

$$\begin{aligned} \theta_i(P) &= \begin{cases} 1 & \text{if } P = N_i, \\ 0 & \text{otherwise} \end{cases} && \text{for all } P \in \mathcal{N}'(\mathcal{T}), \\ \psi_i(P) &= \begin{cases} 1 & \text{if } P = N_i, \\ 0 & \text{otherwise} \end{cases} && \text{for all } P \in \mathcal{N}''(\mathcal{T}). \end{aligned} \tag{9.85}$$

We aim for a conformal discretization, that is, our discretized primal and dual variables $v^h, \phi^h$ will satisfy $v^h \in \mathcal{C}$ and $\phi^h \in \hat{\mathcal{K}}$. Therefore we choose $v^h, (\phi^h)^x \in S^{0,1}(\mathcal{T})$ and $(\phi^h)^s \in S^1(\mathcal{T})$ so that $v^h$ and $\phi^h$ can be written in terms of basis functions as

$$\begin{aligned} v^h(x,s) &= \sum_{k=1}^{q''} V_k \psi_k(x,s), \\ \phi^h(x,s) &= \left( \sum_{k=1}^{q''} \Phi^1_k \psi_k(x,s), \sum_{k=1}^{q''} \Phi^2_k \psi_k(x,s), \sum_{k=1}^{q'} \Phi^s_k \theta_k(x,s) \right), \end{aligned} \tag{9.86}$$

where we denoted the corresponding vectors of nodal function values by capital letters $V, \Phi^1, \Phi^2 \in \mathbf{R}^{q''}$, $\Phi^s \in \mathbf{R}^{q'}$.

**Remark 9.3.1** (Handling of top domain boundary). In the continuous saddle point problem (9.82), the cost functional also includes the integral of the primal and dual function on the top domain boundary $\overline{\Omega} \times \{M\}$; however, we chose to define our discrete functions in $S^{0,1}(\mathcal{T})$ only on $\overline{\Omega} \times [0, M)$. This is unproblematic since in (9.82) we may replace $\overline{\Omega} \times [0, M]$ with $\overline{\Omega} \times [0, M)$ without changing the problem: Since $v = 0$ on $\overline{\Omega} \times [0, M)$ and thus necessarily $D_x v \llcorner \overline{\Omega} \times \{M\} = 0$ and $D_s v \llcorner \overline{\Omega} \times \{M\} \leq 0$, we have

$$\int_{\overline{\Omega} \times \{M\}} \phi \cdot \mathrm{d}Dv = \int_{\overline{\Omega} \times \{M\}} \phi^s \mathrm{d}D_s v \leq 0. \tag{9.87}$$

If this were strictly smaller than zero, then by decreasing $\phi^s$ to zero in a small enough neighborhood of $\overline{\Omega} \times \{M\}$ we could increase $\int_{\overline{\Omega} \times [0,M]} \phi \cdot \mathrm{d}Dv$ so that in the supremum in (9.82) we may indeed ignore the contribution from $\overline{\Omega} \times \{M\}$ without changing its value.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Another way to view this is the observation that $\phi^s(\cdot, M)$ is nothing else but the Lagrange multiplier for the constraint that $v$ must be decreasing in the $s$-direction at $s = M$, which, however, is automatically fulfilled due to the conditions $v \geq 0$ and $v(\cdot, s) = 0$.

Note that an alternative would have been to introduce an auxiliary layer of triangular prism elements right above $\overline{\Omega} \times [0, M)$ so that the discretized functions also have a well-defined value on $\overline{\Omega} \times \{M\}$.

**Remark 9.3.2** (Approximability of the functional). If the triangular prism grid is refined, then we can approximate a continuous function $v$ by discrete functions $v^h$ in the weak-* sense. Note that for a reasonable approximation of functionals involving $Dv$ (as in our case), this is usually not sufficient; instead, we typically need $v^h$ to approximate $v$ in the sense of strict convergence (in which additionally $\|Dv^h\|_{\mathcal{M}} \to \|Dv\|_{\mathcal{M}}$). Unfortunately, this is not possible with a piecewise constant discretization; however, for the special structure of our functional, this would be asking a little bit too much. Indeed, considering for simplicity $v = 1_u$, the cost function satisfies $\mathcal{G}(1_u) = \tilde{\mathcal{E}}(u) = \lim_{n\to\infty} \tilde{\mathcal{E}}(u_n) = \lim_{n\to\infty} \mathcal{G}(1_{u_n})$ for some sequence $u_n$ of piecewise constant images (which follows from Definition 9.2.5 of $\mathcal{E}(\mathcal{F}_u) = \tilde{\mathcal{E}}(u)$ as the relaxation of the cost for discrete mass fluxes). Thus $\mathcal{G}$ can be well approximated even with a discretization that is piecewise constant on a triangular prism grid $\mathcal{T}$ in the $s$-direction. The $x$-derivative of $v$ has to be better resolved, though, to be able to correctly account for the lengths of all network branches. This means that we require the strict convergence in the $x$-direction, $\|D_x v^h\|_{\mathcal{M}} \to \|D_x v\|_{\mathcal{M}}$, and this is indeed ensured by our piecewise linear discretization in the $x$-direction. The discretization of the fluxes $\phi$ now is dual to that of $v$ in the sense that the divergence of $\phi^h$ is also piecewise constant in the $s$-direction and piecewise linear in the $x$-direction. Thus it turns out that from the point of view of the underlying functional lifting, the proposed discretization is a quite natural, conformal one.

Based on this finite element discretization, we can now reformulate the convex saddle point problem (9.82) in terms of the coefficient vectors $V$, $\Phi^1$, $\Phi^2$, $\Phi^s$ as

$$\min_{V \in \mathcal{C}^h(\mathcal{T})} \max_{(\Phi^1, \Phi^2, \Phi^s) \in \hat{\mathcal{K}}^h(\mathcal{T})} V \cdot M^1 \Phi^1 + V \cdot M^2 \Phi^2 + V \cdot M^s \Phi^s, \tag{9.88}$$

where $\mathcal{C}^h(\mathcal{T})$ and $\hat{\mathcal{K}}^h(\mathcal{T})$ are the sets of coefficient vectors corresponding to all functions in $\mathcal{C} \cap S^{0,1}(\mathcal{T})$ and $\hat{\mathcal{K}} \cap (S^{0,1}(\mathcal{T}) \times S^{0,1}(\mathcal{T}) \times S^1(\mathcal{T}))$, respectively, and where $M^1$, $M^2$, $M^s$ denote the mixed mass-stiffness matrices

$$\begin{aligned} M^1_{kl} &= \int_{\overline{\Omega}\times[0,M)} \psi_l \frac{\partial \psi_k}{\partial x_1} \,\mathrm{d}x\,\mathrm{d}s, \quad M^2_{kl} = \int_{\overline{\Omega}\times[0,M)} \psi_l \frac{\partial \psi_k}{\partial x_2} \,\mathrm{d}x\,\mathrm{d}s, \\ M^s_{kl} &= \int_{\overline{\Omega}\times[0,M)} \psi_l \frac{\partial \theta_k}{\partial s} \,\mathrm{d}x\,\mathrm{d}s. \end{aligned} \tag{9.89}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

To explicitly express $\mathcal{C}^h(\mathcal{T})$ and $\hat{\mathcal{K}}^h(\mathcal{T})$, we abbreviate

$$L_{x,s^1,s^2} = \{(x,s) \in \mathcal{N}'(\mathcal{T}) \mid s^1 \le s < s^2\} \tag{9.90}$$

to be the nonhanging nodes with $x$-coordinate $x$ and $s$-coordinate between $s^1$ and $s^2$. Then we can write

$$\begin{aligned}\mathcal{C}^h(\mathcal{T}) = \{&V \in [0,1]^{q''} \mid V_k = 1_{u(\mu_+,\mu_-)}(N_k) \text{ for all } k \in \{1,\dots,q''\} \\ &\text{with } N_k \in \mathcal{N}''(\mathcal{T}) \cap \partial(\Omega \times (0,M))\},\end{aligned} \tag{9.91}$$

$$\begin{aligned}\hat{\mathcal{K}}^h(\mathcal{T}) = \Big\{&(\Phi^1,\Phi^2,\Phi^s) \in (\mathbf{R}^{q''})^2 \times \mathbf{R}^{q'} \;\Big|\; \Phi^s_k \ge 0 \;\forall k = 1,\dots,q', \\ &\Big|\sum_{N_k \in L_{x,s^1,s^2}} h(N_k)(\Phi^1_k,\Phi^2_k)\Big| \le \tau(|s^2 - s^1|) \;\forall\, (x,s^1),(x,s^2) \in \mathcal{N}'(\mathcal{T})\Big\},\end{aligned} \tag{9.92}$$

where $h(N)$ is the distance of $N \in L_{x,s^1,x^2}$ to the next higher node in $L_{x,s^1,x^2}$.

## 9.3.4 Optimization algorithm

We apply an iterative optimization routine that starts on a low-resolution triangular prism grid $\mathcal{T}_0$, on which it solves for the discrete primal and dual variables, resulting in discrete solutions $v_0^h \in S^{0,1}(\mathcal{T}_0)$ and $\phi_0^h \in S^{0,1}(\mathcal{T}_0) \times S^{0,1}(\mathcal{T}_0) \times S^1(\mathcal{T}_0)$. According to some refinement criterion (to be discussed in Section 9.3.5), we then refine several elements of $\mathcal{T}_0$, resulting in a finer grid $\mathcal{T}_1$. On this finer grid we again solve for the discrete primal and dual variables, resulting in $v_1^h$, $\phi_1^h$. We then continue iteratively refining and solving on the grid, thereby producing a hierarchy $\mathcal{T}_0, \mathcal{T}_1, \dots$ of grids with associated discrete solutions $v_k^h$, $\phi_k^h$, $k = 1,2,\dots$.

To solve the discrete saddle point problem on a given grid $\mathcal{T}_k$, we apply a standard primal-dual algorithm [14], in which we perform the projection onto the convex set $\hat{\mathcal{K}}^h(\mathcal{T}_k)$ via an iterative Dykstra routine [5]. This projection is the computational bottleneck of the method (in terms of computation time and memory requirements), and it is the main reason for using the tailored adaptive discretization introduced before. In particular, note that the set of constraints in $\hat{\mathcal{K}}^h(\mathcal{T})$ decomposes into subsets of constraints onto which the projection can be performed independently. In detail, let $x^1,\dots,x^p \in \mathbf{R}^2$ be the nodes of the two-dimensional simplex grid underlying the triangular prism grid and write

$$(\Phi^1,\Phi^2) = ((\Phi^1,\Phi^2)_{x^1},\dots,(\Phi^1,\Phi^2)_{x^p}) \tag{9.93}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

for $(\Phi^1, \Phi^2)_{x^i} = (\Phi^1_k, \Phi^2_k)_{k \in \hat{L}_{x,0,M}}$ and the set $\hat{L}_{x^i,s^1,s^2} = \{k \in \{1, \dots, q''\} \mid N_k \in L_{x,s^1,s^1}\}$ of node indices belonging to $L_{x,s^1,s^2}$. Then

$$\hat{\mathcal{K}}^h(\mathcal{T}) = \left( \bigtimes_{i=1}^{p} \mathcal{K}_{x_i} \right) \times \mathcal{K}_s \tag{9.94}$$

for the convex sets

$$\mathcal{K}_x = \left\{ (\Phi^1_k, \Phi^2_k)_{k \in \hat{L}_{x,0,M}} \;\middle|\; \left| \sum_{k \in \hat{L}_{x,s^1,s^2}} h(N_k)(\Phi^1_k, \Phi^2_k) \right| \leq \tau(|s^2 - s^1|) \; \forall \, (x, s^1), (x, s^2) \in L_{x,0,M} \right\}, \tag{9.95}$$

$$\mathcal{K}_s = \{\Phi^s \in \mathbf{R}^{q'} \mid \Phi^s_k \geq 0 \; \forall k\}. \tag{9.96}$$

so that we can project onto each $\mathcal{K}_{x^i}$ and $\mathcal{K}_s$ separately (where the projection onto $\mathcal{K}_s$ is trivial, and the projection onto each $\mathcal{K}_{x^i}$ is done via Dykstra's algorithm). Note that this would change completely in the presence of $x$-hanging nodes. Here the set $x^1, \dots, x^p$ of simplex grid nodes would also have to include the hanging nodes, and, as a consequence, $\hat{\mathcal{K}}^h(\mathcal{T})$ no longer decomposes into a Cartesion product of constraint sets $\mathcal{K}_{x^i}$, so that the projections can no longer be performed independently.

The overall procedure is presented in pseudocode in Algorithm 9.1 using time steps $\tau, \sigma > 0$ and an overrelaxation parameter $\theta$ from [14] (throughout our numerical experiments, we use $\theta = 1$ and $\tau = \sigma = \frac{1}{L}$ for the Frobenius norm $L$ of the matrix $(M^1, M^2, M^s)$).

## 9.3.5 Refinement criteria

To decide which elements should be refined during the grid refinement in Algorithm 9.1, we use a combination (in our experiments, the maximum) of two heuristic criteria, which both seem to work reasonably well. We define for each element $T \in \mathcal{T}_k$ a refinement indicator $\eta_T(v^h_k, \phi^h_k)$, depending on the solution $(v^h_k, \phi^h_k)$ of the discrete saddle point problem, and we refine any element $T \in \mathcal{T}_k$ with

$$\eta_T(v^h_k, \phi^h_k) \geq \lambda \max_{S \in \mathcal{T}_k} \eta_S(v^h_k, \phi^h_k) \tag{9.97}$$

for some fixed $\lambda \in (0, 1)$.

The first choice of $\eta_T$ is based on the natural and intuitive idea to refine all those elements where the local gradient of the three-dimensional solution $v^h_k$ is high. Indeed, $v^h_k$ approximates a continuous solution, which we expect to be a characteristic function $1_u$, so that by finely resolving regions with high gradient $Dv^h_k$ we expect to

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Algorithm 9.1** Adaptive primal–dual algorithm for generalized branched transport problems.

**function** OPTIMALTRANSPORTNETWORKFE($u^{\mathtt{start}}$,$\mathcal{T}_0$,$\tau$,$\sigma$,$\theta$,numRefinements)
  **for** $i = 0, \dots,$ numRefinements **do**
    assemble matrix $M = (M^1, M^2, M^s)$
    **if** $i = 0$ **then**
      $V^{0,0} = (1_{u^{\mathtt{start}}}(N_1), \dots, 1_{u^{\mathtt{start}}}(N_{q''}))$, $\Psi^{0,0} \equiv (\Phi^1, \Phi^2, \Phi^s)^{0,0} = 0$
    **else**
      prolongate $(V^{i-1,\mathsf{end}}, \Psi^{i-1,\mathsf{end}})$ on $\mathcal{T}_{i-1}$ to $(V^{i,0}, \Psi^{i,0})$ on $\mathcal{T}_i$
    **end if**
    $k \leftarrow 0$
    **while** not converged **do**
      $\tilde{\Psi}^{i,k+1} = \tilde{\Psi}^{i,k} + \sigma M^* \bar{V}^{i,k}$
      compute the projection $\Psi^{i,k+1} = \pi_{\hat{\mathcal{K}}^h}(\tilde{\Psi}^{i,k+1})$ via separate projections onto sets $\mathcal{K}_{x^i}, \mathcal{K}_s$ using Dykstra's algorithm
      $\tilde{V}^{i,k+1} = V^{i,k} - \tau M \Psi^{i,k+1}$
      compute the projection $V^{i,k+1} = \pi_{\mathcal{C}^h}(\tilde{V}^{i,k+1})$
      $\bar{V}^{i,k+1} = V^{i,k+1} + \theta(V^{i,k+1} - V^{i,k})$
      $k \leftarrow k + 1$
    **end while**
    **if** $i <$ numRefinements **then**
      refine grid $\mathcal{T}_i$ to $\mathcal{T}_{i+1}$
    **else**
      $V = V^{i,\mathsf{end}}$, $(\Phi^1, \Phi^2, \Phi^s) = \Psi^{i,\mathsf{end}}$
    **end if**
  **end for**
**end function**
**return** $V, \Phi^1, \Phi^2, \Phi^s$

better approximate $1_u$. Thus we define

$$\eta_T(v_k^h, \phi_k^h) = \frac{1}{\mathcal{L}^3(T)} |Dv_k^h|(T'). \tag{9.98}$$

Although this strategy is computationally cheap and easy to handle, gradient refinement only takes the current grid structure into account and neglects any information about the functional (possibly leading to redundantly refined elements).

The second choice of $\eta_T$ is (an approximation of) the local primal–dual gap, that is, the contribution of each element to the global primal–dual gap

$$\Delta(v_k^h, \phi_k^h) = \mathcal{G}(v_k^h) - \mathcal{D}(\phi_k^h) \geq 0 \tag{9.99}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

associated with the strong duality from Theorem 9.2.3. Since $\Delta(v_k^h, \phi_k^h) = 0$ implies that $(v_k^h, \phi_k^h)$ is the global solution of the saddle point problem, it is natural to refine the grid in those regions where the largest contribution to the duality gap occurs. This contribution can be calculated as follows:

$$\begin{aligned}\Delta(v_k^h, \phi_k^h) &= \sup_{\phi \in \hat{\mathcal{K}}} \int_{\overline{\Omega}\times[0,M)} \phi \cdot \mathrm{d}Dv_k^h - \min_{v\in\mathcal{C}} \int_{\overline{\Omega}\times[0,M)} \phi_k^h \cdot \mathrm{d}Dv \\ &= \sup_{\phi \in \hat{\mathcal{K}}} \int_{\overline{\Omega}\times[0,M)} (\phi - \phi_k^h) \cdot Dv_k^h \,\mathrm{d}x\,\mathrm{d}s - \min_{v\in\mathcal{C}} \int_{\overline{\Omega}\times[0,M)} \phi_k^h \cdot \mathrm{d}D(v - v_k^h) \\ &= \sup_{\phi \in \hat{\mathcal{K}}} \int_{\overline{\Omega}\times[0,M)} (\phi^x - (\phi_k^h)^x) \cdot D_x v_k^h \,\mathrm{d}x\,\mathrm{d}s \\ &\quad + \max_{v\in\mathcal{C}} \int_{\overline{\Omega}\times[0,M)} (v - v_k^h) \operatorname{div} \phi_k^h \,\mathrm{d}x\,\mathrm{d}s. \end{aligned} \tag{9.100}$$

Although the maximizing $v^{\mathrm{opt}}$ can readily be calculated as $v^{\mathrm{opt}}(x,s) = \max\{0, \operatorname{sign}(\operatorname{div}\phi_k^h)\}$, the supremum has no analytical expression and needs to be evaluated numerically. We approximate it by refining $\mathcal{T}_k$ uniformly to some grid $\tilde{\mathcal{T}}_k$ and then calculating

$$\phi^{\mathrm{opt}} = \operatorname*{arg\,max}_{\phi \in \hat{\mathcal{K}}^h(\tilde{\mathcal{T}}_k)} \int_{\overline{\Omega}\times[0,M)} \phi^x \cdot D_x v_k^h \,\mathrm{d}x\,\mathrm{d}s \in S^{0,1}(\mathcal{T}_k) \times S^{0,1}(\mathcal{T}_k) \times \{0\}. \tag{9.101}$$

Note that this latter maximization can be independently performed for the function values at nodes with different $x$-coordinates and thus is very fast. We then set the refinement indicator as

$$\eta_T(v_k^h, \phi_k^h) = \int_T ((\phi^{\mathrm{opt}})^x - (\phi_k^h)^x) \cdot D_x v_k^h \,\mathrm{d}x\,\mathrm{d}s + \int_T (v^{\mathrm{opt}} - v_k^h) \operatorname{div} \phi_k^h \,\mathrm{d}x\,\mathrm{d}s. \tag{9.102}$$

Since $\phi^{\mathrm{opt}}$ is only an approximation of the true minimizer, $\sum_{T\in\mathcal{T}_k} \eta_T(v_k^h, \phi_k^h) \le \Delta(v_k^h, \phi_k^h)$ is an approximation of the duality gap from below. Note that the summand $\int_T (v^{\mathrm{opt}} - v_k^h) \operatorname{div} \phi_k^h \,\mathrm{d}x\,\mathrm{d}s$ is nonnegative, whereas $\int_T ((\phi^{\mathrm{opt}})^x - (\phi_k^h)^x) \cdot D_x v_k^h \,\mathrm{d}x\,\mathrm{d}s$ in principle may have either sign. However, at least, we have $\sum_{x^i \in \pi_x(T)} \int_T ((\phi^{\mathrm{opt}})^x - (\phi_k^h)^x) \cdot D_x v_k^h \,\mathrm{d}x\,\mathrm{d}s \ge 0$ for all simplex grid nodes $x^i$ (where $\pi_x : \mathbf{R}^3 \to \mathbf{R}^2$ is the projection onto the first two coordinates), so that $\eta_T(v_k^h, \phi_k^h)$ may well serve as a local refinement indicator.

### 9.3.6 Results

We implemented the algorithm described above in C++, where the grid and corresponding finite element classes are based on the QuocMesh library [33]. For our ex-

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

periments, we pick the branched transport and urban planning transportation costs $\tau$ from Example 9.2.1.

To begin with, we test the reliability of the method by comparing its results with the true solution in a simple symmetric setting in which the optimal transport network can actually be calculated by hand. This setting has four evenly spaced point sources of equal mass at the top side of the rectangular domain $\Omega = [0,1]^2$ and four evenly spaced point sinks of same mass exactly opposite. Due to the high symmetry, there are only a handful of possible graph topologies whose vertex positions can explicitly be optimized. For both branched transport and urban planning, we test a range of parameters to explore multiple different topologies. Figures 9.7 and 9.8 show that in each case the algorithm converged to the correct solution except for one parameter setting close to a bifurcation point where the optimal network topology changes. In that setting, our algorithm returned a convex combination of functions $1_u$ corresponding to two different topologies, which numerically both seem to be of sufficiently equal optimality so that the algorithm converges to their convex combination (compare Propo-

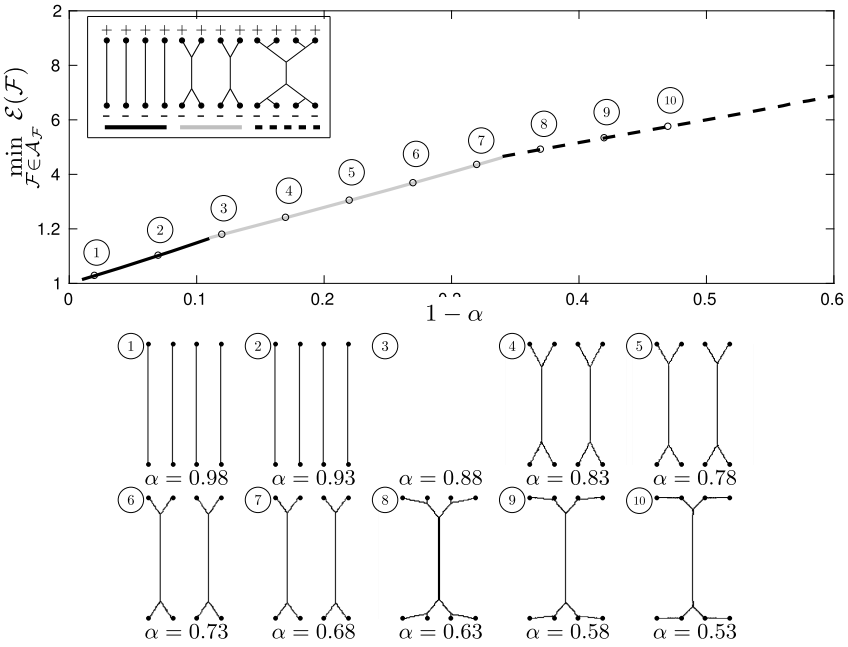



**Figure 9.7:** Parameter study for branched transport with transportation cost $\tau^{\mathrm{bt}}(m) = m^\alpha$. Top: Plot of the manually and numerically computed minimal energy for different values of $1-\alpha$. The line type indicates the optimal network topology. Bottom: Numerically computed optimal transport networks for evenly spaced values of $\alpha$ in the same range (if the numerical solution is $v$, then we show the support of the gradient of its projection onto the $x$-plane). The numerically obtained network topologies match the predicted ones except for Example ③, where the three-dimensional solution is not binary but a convex combination of the binary solutions to two different topologies.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

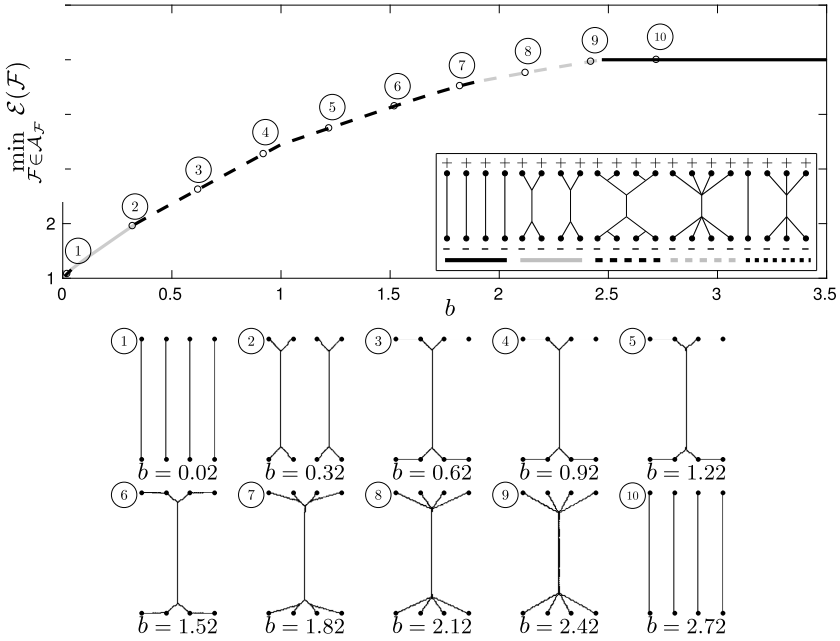



**Figure 9.8:** Parameter study for urban planning with transportation cost $\tau^{\mathrm{up}}(m) = \min\{am, m + b\}$ for $a = 5$ and varying $b$. Illustration as in Figure 9.7.

sition 9.2.1(3)). This is in fact a slight improvement over the result in [7], where we performed exactly the same experiment, only using a standard finite difference discretization at much lower resolution. For that discretization and resolution, the algorithm actually converged to the wrong topology, which was better aligned with the grid and therefore advantageous at the given resolution. With our new discretization, we achieve a higher resolution, enabling the algorithm to move away from that erroneous topology. It seems that more grid refinement would be necessary to recover the true solution; however, to make the results for all parameters comparable, we chose the same number of refinements throughout Figures 9.7 and 9.8.

Note that the reliability of the algorithm is not obvious a priori since an adaptive refinement may in principle lead to discretization artefacts, giving preference to material fluxes through highly resolved areas over fluxes through coarsly discretized areas, in which the discretization error produces artificial additional costs.

At this point, we would also like to mention that in [7], we obtained one simulation result for urban planning with the same four sources and sinks as in Figure 9.8 (but different parameter values), which was not binary, and which we assumed to be a manifestation of the convex relaxation being not tight. However, it turns out that the result was again just a convex combination of two global minimizers, namely the

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

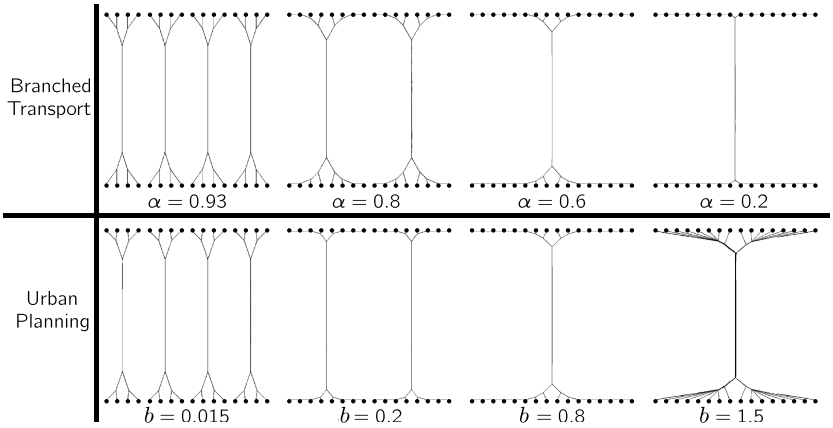



**Figure 9.9:** Numerical optimization results for transport from 16 almost evenly spaced point sources to 16 point sinks of the same mass ($a = 5$ for the urban planning results).

right-most topology in Figure 9.8 and its mirror image (which just happen to be never optimal for the parameters in Figure 9.8).

Next, we repeat the other numerical simulations from [7], which require transport networks of much more complex branching structure, and which due to a lack of resolution, could hardly be resolved in [7] (in fact, the smallest obtained network branches were at the order of the discretization width, and all network branches were visibly distorted by the pixel grid). Figures 9.9 and 9.10 show simulation results for these configurations with much more satisfying accuracy at which all branches are clearly resolved. In these rather symmetric example settings, we slightly broke the symmetry by perturbing the even spacing of sources and sinks, since otherwise there would be multiple global optimal transport networks, a convex combination of which would be returned by our algorithm. This symmetry break might also be the reason why the challenging simulation of branched transport for $\alpha = 0.99$ exhibits faintly Y-shaped branches everywhere except for the rightmost four branches, which lead straight from the source to the sink (however, as an alternative explanation, our final level of mesh adaptivity might simply still be too coarse to resolve Y-shaped branches along those particular directions due to local mesh anisotropies).

To be able to have a source point within the domain $\Omega$ in Figure 9.10 (recall that $\mu_+, \mu_-$ should lie on $\partial\Omega$), we employ the following trick: we connect the center source with boundary $\partial\Omega$ by a (straight) line across which we enforce the variables $v$ and $\phi$ to be discontinuous with

$$v^-(x,s) = v^+(x,s+M), \quad \phi^-(x,s) = \phi^+(x,s+M) \tag{9.103}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

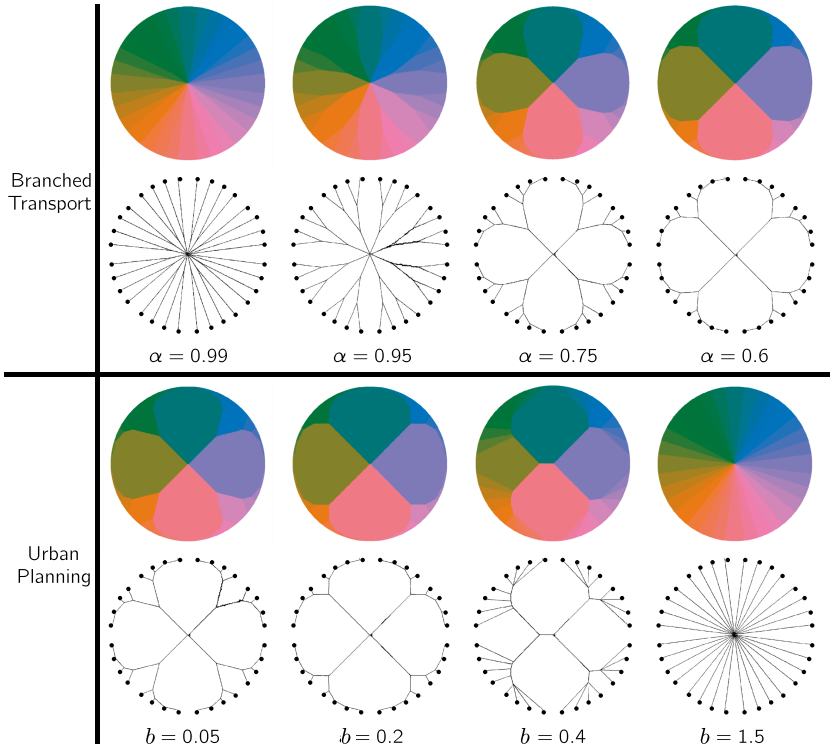



**Figure 9.10:** Numerical optimization results for transport from a central source point to 32 almost evenly spaced point sinks of equal mass on a concentric circle ($a = 5$ for the urban planning results). Using a periodic color-coding, we show the images $u$, whose lifting $1_u$ is the numerical solution, and the support of their gradient underneath, which represents the transport network.

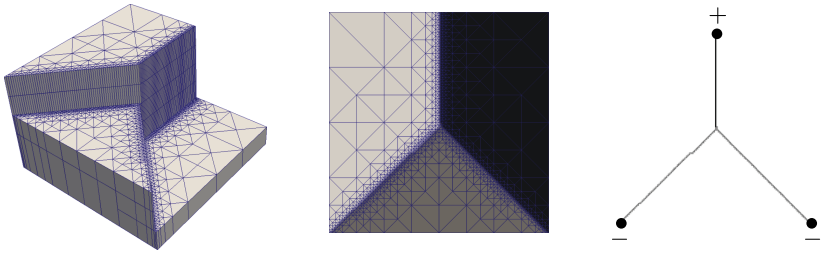


**Figure 9.11:** Optimal network for branched transport from one point mass at the top of $\Omega = [0,1]^2$ to two equal sinks at the bottom corners of $\Omega$. Left: Profile of the three-dimensional discrete solution $v^h$ (the displayed surface shows the $\frac{1}{2}$-level set of $v$ with finite element boundaries indicated in blue). Middle: Two-dimensional image obtained by projecting $v^h$ onto the $x$-plane (the element boundaries of the underlying two-dimensional simplex grid are shown in blue). Right: Optimal network structure given by the support of the image gradient.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

for $M = \|\mu_+\|_{\mathcal{M}} = \|\mu_-\|_{\mathcal{M}}$. Essentially, this means that we take the range of the two-dimensional images $u$ (corresponding to the mass fluxes) to be an infinite covering of $[0, M)$ with fibers $r + M\mathbb{Z}$.

We finally discuss the gain in computational efficiency by the new adaptive discretization. We already saw before that the adaptive discretization allows us to produce a quality of the transport networks that goes far beyond a standard discretization. At the same time the computational cost decreases. Figure 9.11 illustrates, for a simple example that can readily be visualized, the reason for the enhanced efficiency, the underlying adaptive grid refinement near the network branches. Table 9.2 and Figure 9.12 quantify the speedup of going from a standard uniform discretization to the

**Table 9.2:** Comparison between branched transport network simulations on uniform and adaptive grids. The first column refers to the $x$- and $s$-levels of the uniform grid and the highest local $x$- and $s$-levels of the adaptive grid (the $x$- and $s$-levels of an element are the numbers of $x$- and $s$-bisections necessary to obtain the element starting from an element of the same size as the computational domain). The table shows the number of elements, of degrees of freedom in the variable $v^h$, the runtime, and the calculated primal–dual gap at the end. For the adaptive simulation, the relative number of elements and degrees of freedom compared to the uniform simulation is also shown as a percentage. All adaptive simulations start at a uniform grid of $x$-level 4 and $s$-level 2. The experiments on a uniform grid of the highest levels are omitted due to their infeasible runtime and memory consumption.

| x/s | Uniform | | | | Adaptive | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | numEls | numDofs | time | pd gap | numEls | numDofs | %Els | %Dofs | time | pd gap |
| 4/2 | 2048 | 1445 | 14 sec. | 0.0069 | 2048 | 1445 | 100 | 100 | 14 sec. | 0.0069 |
| 5/3 | 16384 | 9801 | 96 sec. | 0.0192 | 7111 | 4576 | 43.4 | 46.7 | 44 sec. | 0.0101 |
| 6/4 | 131072 | 71825 | 855 sec. | 0.0165 | 30961 | 18800 | 23.6 | 26.2 | 184 sec. | 0.0431 |
| 7/5 | 1048576 | 549153 | 20014 sec. | 0.0013 | 91391 | 53596 | 8.7 | 9.8 | 632 sec. | 0.0027 |
| 8/6 | 8388608 | 4293185 | 224221 sec. | 0.0047 | 146825 | 84749 | 1.7 | 2.0 | 1405 sec. | 0.0019 |
| 9/7 | – | – | – | – | 295227 | 167030 | 0.4 | 0.5 | 3438 sec. | 0.0008 |
| 10/8 | – | – | – | – | 667289 | 370570 | 0.1 | 0.1 | 9767 sec. | 0.0003 |

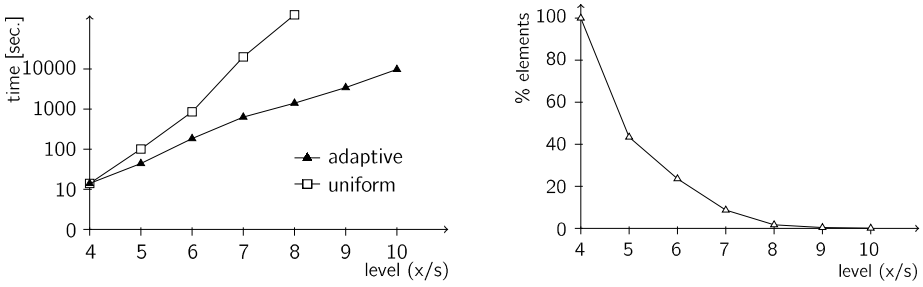



**Figure 9.12:** Runtime and relative number of elements in a simulation on an adaptive versus a uniform grid from Table 9.2.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

adaptive one (for the same configuration as in Figure 9.7 with $\alpha = 0.5$), which quickly reaches orders of magnitude.

## 9.4 Discussion

We shed more light on the relation between two-dimensional generalized branched transport and corresponding convex optimization problems obtained via functional lifting. In particular, it is now clear that those problems are indeed equivalent up to a relaxation step whose tightness is expected but not known. With a tailored adaptive finite element discretization, this relation can now be leveraged to solve two-dimensional generalized branched transport problems.

A seeming disadvantage of the functional lifting approach lies in the fact that the given material source and sink $\mu_+$, $\mu_-$ need to be supported on the computational domain boundary. This deficiency can be overcome by a trick similar to that of Figure 9.10, introduced in [3]. To this end, we fix an initial backward mass flux $\mathcal{F}_-$ from $\mu_-$ to $\mu_+$. Taking now any mass flux $\mathcal{F}$ from $\mu_+$ to $\mu_-$, the joint flux $\mathcal{F} + \mathcal{F}_-$ has zero divergence and can thus be translated into the gradient of an image. During the image optimization or the corresponding lifted convex optimization, we just have to ensure by constraints that the backward mass flux stays fixed and is not changed (and also we have to adapt the cost functional so as to neglect the cost of $\mathcal{F}_-$ and to prevent artificial cost savings that may come about by aggregating part of $\mathcal{F}$ with $\mathcal{F}_-$).

A true disadvantage, though, of the approach is that it is inherently limited to two space dimensions. Indeed, it exploits that in two space dimensions the one-dimensional network structures also have codimension 1 and thus can be interpreted as image gradients. However, the two-dimensional case is of importance in various settings such as logistic problems, public transport networks, river networks or leaf venation, to name but a few examples.

Compared to graph-based methods, computation times of our approach are of course much longer; however, our approach is guaranteed to yield an approximation of a global minimizer. Indeed, even though here we did not show the $\Gamma$-convergence of the discretized functional to the continuous one (which together with the global BV-coercivity implies the weak-* convergence of discrete to continuous optima as the computational grid is refined), the experienced reader will notice that in essence it will follow in the same way as for finite element discretizations of the standard BV-seminorm. Nevertheless, heuristic topology optimization procedures on graphs seem to result in networks of almost the same quality. It is conceivable that a combination of both approaches may increase efficiency while maintaining the guarantee of a global minimum.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

# Bibliography


[1] G. Alberti, G. Bouchitté, and G. Dal Maso. The calibration method for the Mumford–Shah functional and free-discontinuity problems. *Calc. Var. Partial Differ. Equ.*, 16(3):299–333, 2003.

[2] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford Science Publications. Clarendon Press, 2000.

[3] M. Bonafini, G. Orlandi, and É. Oudet. Variational approximation of functionals defined on 1-dimensional connected sets: the planar case. *SIAM J. Math. Anal.*, 50(6):6307–6332, 2018.

[4] J. Borwein and Q. Zhu. *Techniques of Variational Analysis*. Springer, New York, 2005.

[5] J. P. Boyle and R. L. Dykstra. A method for finding projections onto the intersection of convex sets in Hilbert spaces. In *Advances in Order Restricted Statistical Inference*, volume 37 of Lecture Notes in Statistics, pages 28–47. Springer, New York, 1986.

[6] A. Brancolini and G. Buttazzo. Optimal networks for mass transportation problems. *ESAIM Control Optim. Calc. Var.*, 11(1):88–101, 2005.

[7] A. Brancolini, C. Rossmanith, and B. Wirth. Optimal micropatterns in 2D transport networks and their relation to image inpainting. *Arch. Ration. Mech. Anal.*, 228(1):279–308, Apr 2018.

[8] A. Brancolini and B. Wirth. Equivalent formulations for the branched transport and urban planning problems. *J. Math. Pures Appl.*, 106(4):695–724, 2016.

[9] A. Brancolini and B. Wirth. General transport problems with branched minimizers as functionals of 1-currents with prescribed boundary. *Calc. Var. Partial Differ. Equ.*, 57(3):Art. 82, 39, 2018.

[10] A. Bressan and Q. Sun. On the optimal shape of tree roots and branches. *Math. Models Methods Appl. Sci.*, 28(14):2763–2801, 2018.

[11] G. Buttazzo, A. Pratelli, S. Solimini, and E. Stepanov. *Optimal Urban Networks via Mass Transportation*, volume 1961 of Lecture Notes in Mathematics. Springer-Verlag, Berlin, 2009.

[12] A. Chambolle, B. Merlet, and L. Ferrari. A simple phase-field approximation of the Steiner problem in dimension two. *Adv. Calc. Var.*, 12(2):157–179, 2019.

[13] A. Chambolle, B. Merlet, and L. Ferrari. Strong approximation in $h$-mass of rectifiable currents under homological constraint. *Adv. Calc. Var.*, 14(3):343–363, 2021.

[14] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40(1):120–145, 2011.

[15] M. Fampa, J. Lee, and N. Maculan. An overview of exact algorithms for the Euclidean Steiner tree problem in $n$-space. *Int. Trans. Oper. Res.*, 23(5):861–874, 2016.

[16] L. Ferrari, C. Rossmanith, and B. Wirth. Phase field approximations of branched transportation problems. Preprint, 2019.

[17] L. A. D. Ferrari. Phase-field approximation for some branched transportation problems. Theses, Université Paris-Saclay, October 2018.

[18] R. Fonseca, M. Brazil, P. Winter, and M. Zachariasen. Faster exact algorithm for computing Steiner trees in higher dimensional Euclidean spaces. In *11th DIMACS Implementation Challenge Workshop*, Providence, RI, 2014. http://dimacs11.cs.princeton.edu/workshop/FonsecaBrazilWinterZachariasen.pdf.

[19] E. N. Gilbert. Minimum cost communication networks. *Bell Syst. Tech. J.*, 46(9):2209–2227, Nov 1967.

[20] E. N. Gilbert and H. O. Pollak. Steiner minimal trees. *SIAM J. Appl. Math.*, 16:1–29, 1968.

[21] D. Juhl, D. M. Warme, P. Winter, and M. Zachariasen. The GeoSteiner software package for computing Steiner trees in the plane: an updated computational study. *Math. Program. Comput.*, 10(4):487–532, 2018.

[22] N. Maculan, P. Michelon, and A. Xavier. The Euclidean Steiner tree problem in $\mathbb{R}^n$: a mathematical programming formulation. *Ann. Oper. Res.*, 96(1):209–220, 2000.


EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[23] F. Maddalena, S. Solimini, and J.-M. Morel. A variational model of irrigation patterns. *Interfaces Free Bound.*, 5:391–415, 12 2003.
[24] A. Marchese and B. Wirth. Approximation of rectifiable 1-currents and weak-$*$ relaxation of the $h$-mass. *J. Math. Anal. Appl.*, 479(2):2268–2283, 2019.
[25] M. Matuszak, J. Miekisz, and T. Schreiber. Solving ramified optimal transport problems in the Bayesian influence diagram framework. In *International Conference on Artificial Intelligence and Soft Computing*. Lecture Notes in Computer Science, pages 582–590, 2012.
[26] Z. A. Melzak. On the problem of Steiner. *Can. Math. Bull.*, 4(2):143–148, 1961.
[27] A. Monteil. Uniform estimates for a Modica–Mortola type approximation of branched transportation. *ESAIM Control Optim. Calc. Var.*, 23(1):309–335, 2017.
[28] D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems. *Commun. Pure Appl. Math.*, 17:577–685, 1989.
[29] E. Oudet and F. Santambrogio. A Modica–Mortola approximation for branched transport and applications. *Arch. Ration. Mech. Anal.*, 201(1):115–142, 2011.
[30] J. Piersa. Ramification algorithm for transporting routes in $\mathbb{R}^2$. In *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*, Limassol, pages 657–664, 2014.
[31] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. An algorithm for minimizing the Mumford–Shah functional. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1133–1140, Sep 2009.
[32] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. Global solutions of variational models with convex regularization. *SIAM J. Imaging Sci.*, 3(4):1122–1145, Dec 2010.
[33] QuocMesh. Using and programming the QuocMesh library. QuocMesh Collective. Version 1.5, 2014. https://archive.ins.uni-bonn.de/numod.ins.uni-bonn.de/software/quocmesh/1.5/doc/lib/index.html.
[34] S. K. Smirnov. Decomposition of solenoidal vector charges into elementary solenoids, and the structure of normal one-dimensional flows. *Algebra Anal.*, 5(4):206–238, 1993.
[35] W. D. Smith. How to find Steiner minimal trees in Euclidean $d$-space. *Algorithmica*, 7(1–6):137–177, 1992.
[36] C. T. Traxler. An algorithm for adaptive mesh refinement in $n$ dimensions. *Computing*, 59:115–137, 1997.
[37] B. Wirth. Phase field models for two-dimensional branched transportation problems. *Calc. Var. Partial Differ. Equ.*, 58(5):Art. 164, 31, 2019.
[38] Q. Xia. Optimal paths related to transport problems. *Commun. Contemp. Math.*, 5(2):251–279, 2003.
[39] Q. Xia. The formation of a tree leaf. *ESAIM Control Optim. Calc. Var.*, 13(2):359–377, 2007.
[40] Q. Xia. Numerical simulation of optimal transport paths. In *2010 Second International Conference on Computer Modeling and Simulation*, 2008.
[41] Q. Xia. Motivations, ideas and applications of ramified optimal transportation. *ESAIM: Math. Model. Numer. Anal.*, 49(6):1791–1832, 2015.
[42] Q. Xia, L. A. Croen, M. Danielle Fallin, C. J. Newschaffer, C. Walker, P. Katzman, R. K. Miller, J. Moye, S. Morgan, and C. Salafia. Human placentas, optimal transportation and high-risk autism pregnancies. *J. Coupled Syst. Multiscale Dyn.*, 4(4):260–270, 2016.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Florian Feppon

# 10 High-order homogenization of the Poisson equation in a perforated periodic domain

**Abstract:** We derive high-order homogenized models for the Poisson problem in a cubic domain periodically perforated with holes, where Dirichlet boundary conditions are applied. These models have the potential to unify three possible kinds of limit problems derived in the literature for various asymptotic regimes (namely, the "unchanged" Poisson equation, the Poisson problem with a strange reaction term, and the zeroth-order limit problem) of the ratio $\eta \equiv a_\epsilon/\epsilon$ between the size $a_\epsilon$ of the holes and the size $\epsilon$ of the periodic cell. The derivation relies on algebraic manipulations on formal two-scale power series in terms of $\epsilon$ and more particularly on the existence of a "criminal" ansatz, which allows us to reconstruct the oscillating solution $u_\epsilon$ as a linear combination of the derivatives of its formal average $u_\epsilon^*$ weighted by suitable corrector tensors. The formal average is itself a solution of a formal infinite-order homogenized equation. Classically, truncating the infinite-order homogenized equation yields in general an ill-posed model. Inspired by a variational method introduced in [23, 52], we derive, for any $K \in \mathbb{N}$, well-posed corrected homogenized equations of order $2K+2$, which yield approximations of the original solutions with an error of order $O(\epsilon^{2K+4})$ in the $L^2$ norm. Finally, we find asymptotics of all homogenized tensors in the low volume fraction regime $\eta \to 0$ and in dimension $d \geq 3$. This allows us to show that our higher-order effective equations converge coefficientwise to either of the classical homogenized regimes of the literature, which arise when $\eta$ is respectively equivalent to or greater than the critical scaling $\eta_{\text{crit}} \sim \epsilon^{2/(d-2)}$.

**Keywords:** homogenization, higher-order models, perforated Poisson problem, homogeneous Dirichlet boundary conditions, strange term

**MSC 2010:** 35B27, 76M50, 35J30

## 10.1 Introduction

One of the industrial perspectives offered today by the theory of homogenization lies in the development of more and more efficient topology optimization algorithms for the

**Acknowledgement:** This work was supported by the Association Nationale de la Recherche et de la Technologie (ANRT) [grant number CIFRE 2017/0024] and by the project ANR-18-CE40-0013 SHAPO financed by the French Agence Nationale de la Recherche (ANR). The author is grateful to G. Allaire, C. Dapogny and S. Fliss for insightful discussions and their helpful comments and revisions.

**Florian Feppon,** Centre de Mathématiques Appliquées, École Polytechnique, Palaiseau, France; and Safran Tech, Magny-les-Hameaux, France, e-mail: florian.feppon@polytechnique.edu

https://doi.org/10.1515/9783110695984-010

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

design of mechanical structures; there exists a variety of homogenization-based techniques [7, 16, 24], including density-based (or "SIMP") methods [17]. Broadly speaking, the principle of these algorithms is optimizing one or several parameters of the microstructures, which affect the coefficients of an effective model. The latter model accounts for the constitutive physics of the mixture of two materials (one representing the solid structure and one representing void) and is mathematically obtained by homogenization of the linear elasticity system. The knowledge of the dependence between the parameters of the microstructure and the coefficient of the effective model is the key ingredient of homogenization-based techniques, because it allows us to numerically – and automatically – interpret "gray designs" (i. e., for which the local density of solid is not a uniformly equal to 0 or 1) into complex composite shapes characterized by multiscale patterns and geometrically modulated microstructures [11, 30, 36, 45].

Several works have sought extensions of these methods for topology optimization of fluid systems, where the incompressible Navier–Stokes system is involved. In this context, an effective model is needed for describing the homogenized physics of a porous medium filled with either solid obstacles, fluid, or a mixture of both. However, the classical literature [3, 5, 25, 49] identifies three possible kinds of homogenized models depending on how periodic obstacles of size $a_\epsilon$ scale within their periodic cell of size $\epsilon$: depending on how the scaling $\eta \equiv a_\epsilon/\epsilon$ compares to the critical size $\sigma_\epsilon := \epsilon^{d/(d-2)}$ (in dimension $d \geq 3$), the fluid velocity converges as $\epsilon \to 0$ to the solution of either a Darcy, a Brinkman or a Navier–Stokes equation. Unfortunately, there is currently no further result regarding an effective model that would be able to describe a medium featuring all possible sizes of locally periodic obstacles. The strategy that is the most commonly used in the density-based topology optimization community consists in using either the Brinkman equation exclusively [20, 21, 28] or the Darcy model exclusively [47, 54]. These methodologies have proved to be efficient in a number of works [27, 46]; however, they remain inconsistent from a homogenization point of view, since these models are valid only for particular regimes of size of obstacles. In particular, this limitation makes impossible to interpret "gray" designs obtained with classical fluid topology optimization algorithms.

The main objective of this paper is to expose, accordingly, the derivation of a new class of – high-order – homogenized models for perforated problems that have the potential to unify the different regimes of the literature. Our hope is that these models could allow us, in future works, to develop new mathematically consistent and homogenization-based topology optimization algorithms for fluid systems.

This paper is a preliminary study toward such purpose: we propose to investigate the case of the Poisson problem in a perforated periodic domain with Dirichlet bound-

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

ary conditions on the holes,

$$\begin{cases} -\Delta u_\epsilon = f \quad \text{in } D_\epsilon, \\ u_\epsilon = 0 \quad \text{on } \partial\omega_\epsilon, \\ u_\epsilon \text{ is } D\text{-periodic}, \end{cases} \tag{10.1}$$

which can be considered to be a simplified scalar and linear version of the full Navier–Stokes system. Let us mention that the analysis of the full Navier–Stokes system is much more challenging because of the incompressibility constraint and the vectorial nature of the problem; the extension of the current work for the Stokes system – its linear counterpart – will however be exposed in a future contribution [33, 34].

The setting considered is the classical context of periodic homogenization represented on Figure 10.1: $D := [0, L]^d$ is a $d$-dimensional box filled with periodic obstacles $\omega_\epsilon := \epsilon(\mathbb{Z}^d + \eta T) \cap D$. We denote by $P = (0,1)^d$ the unit cell, and by $Y := P \setminus \eta T$ the unit perforated cell. The parameter $\epsilon$ is the size of the periodic cell and is given by $\epsilon := L/n$, where $n \in \mathbb{N}$ is an integer number assumed to be large. The parameter $\eta$ is another rescaling of the obstacle $T$ within the unit cell: the holes are therefore of size $a_\epsilon := \eta\epsilon$, which allow us to consider in Section 10.5 the so-called low volume fraction limits as $\eta$ converges to zero. The boundary of the obstacle $T$ is assumed to be smooth, $D_\epsilon := D \setminus \omega_\epsilon$ denotes the perforated domain, and $f \in C^\infty(D)$ is a smooth $D$-periodic right-hand side. The periodicity assumption for $u_\epsilon$ and $f$ is classical in homogenization and is used to avoid difficulties related to the arising of boundary layers (see [8, 19, 39]).

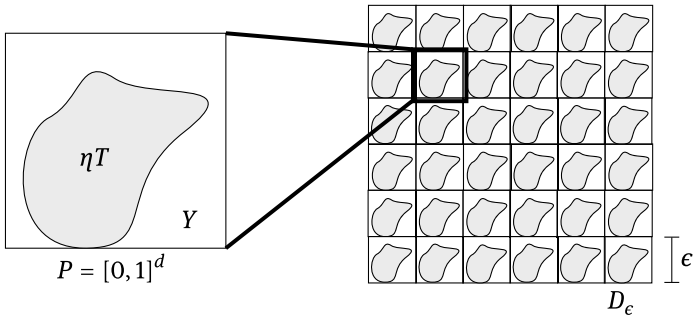



**Figure 10.1:** The perforated domain $D_\epsilon$ and the unit cell $Y = P \setminus (\eta T)$.

The literature accounts for several homogenized equations depending on how the size $a_\epsilon = \eta\epsilon$ of the holes compares to the critical size $\sigma_\epsilon := \epsilon^{d/(d-2)}$ in dimension $d \geq 3$ (or how $\log(a_\epsilon)$ compares to $\log(\sigma_\epsilon) := -1/\epsilon^2$ in dimension $d = 2$) [3, 5, 26, 29, 37, 43, 49]:

– if $a_\epsilon = o(\sigma_\epsilon)$ (if $\log(a_\epsilon) = o(\log(\sigma_\epsilon))$ if $d = 2$), then the holes are "too small", and $u_\epsilon$ converges, as $\epsilon \to 0$, to the solution $u$ of the Poisson equation in the homogeneous

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

domain $D$ (without holes):

$$\begin{cases} -\Delta u = f \quad \text{in } D, \\ u \text{ is } D\text{-periodic.} \end{cases} \tag{10.2}$$

- if $a_\epsilon = \sigma_\epsilon$ (if $\log(a_\epsilon) = -1/\epsilon^2$ if $d = 2$), then $u_\epsilon$ converges, as $\epsilon \to 0$, to the solution $u$ of the modified Poisson equation

$$\begin{cases} -\Delta u + Fu = f \quad \text{in } D, \\ u \text{ is } D\text{-periodic,} \end{cases} \tag{10.3}$$

  where the so-called *strange* reaction term $Fu$ involves a positive constant $F > 0$, which can be computed by means of an exterior problem in $\mathbb{R}^d \setminus T$ when $d \geq 3$ (see equation (10.80) below), and which is equal to $2\pi$ if $d = 2$ (see [3, 25, 37, 49]).
- if $\sigma_\epsilon = o(a_\epsilon)$ (if $\log(\sigma_\epsilon) = o(\log(a_\epsilon))$ if $d = 2$) and $a_\epsilon = \eta\epsilon$ with $\eta \to 0$ as $\epsilon \to 0$, then the holes are "large", and $a_\epsilon^{d-2}\epsilon^{-d}u_\epsilon$ ($-\log(a_\epsilon)\epsilon^{-2}u_\epsilon$ if $d = 2$) converges to the solution $u$ of the zeroth-order equation

$$\begin{cases} Fu = f \quad \text{in } D, \\ u \text{ is } D\text{-periodic,} \end{cases} \tag{10.4}$$

  where $F$ is the same positive constant as in equation (10.3).
- if $a_\epsilon = \eta\epsilon$ with the ratio $\eta$ *fixed*, then $\epsilon^{-2}u_\epsilon$ converges to the solution $u$ of the zeroth-order equation

$$\begin{cases} M^0 u = f \quad \text{in } D, \\ u \text{ is } D\text{-periodic,} \end{cases} \tag{10.5}$$

  where $M^0$ is another positive constant (which depends on $\eta$). Furthermore, it can be shown that $M^0/|\log(\eta)| \to F$ if $d = 2$, and $M^0/\eta^{d-2} \to F$ (if $d \geq 3$) as $\eta \to 0$, so that there is a continuous transition from equation (10.5) to equation (10.4); see [4] and Corollary 10.5.1.

The different regimes (10.2)–(10.5) occur because the heterogeneity of the problem comes from the zero Dirichlet boundary condition on the holes $\omega_\epsilon$ in equation (10.1): this is the major difference with the setting commonly assumed in linear elasticity, where the heterogeneity induced by the mixture of two materials is instead inscribed in the coefficients of the physical state equation [7, 53]. Equations (10.2) and (10.3) are, respectively, the analogs of the Navier–Stokes and Brinkman regimes in the context of the homogenization of the Navier–Stokes equation, whereas the zeroth-order equations (10.4) and (10.5) are analogous to Darcy models. As stressed above, the existence of these regimes raises practical difficulties in view of applying the homog-

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

enization method for shape optimization: the previous considerations show that we should use (10.2) and (10.3) in regions featuring none or very tiny obstacles; however, we should use the zeroth-order model (10.5) when the obstacles become large enough.

Our goal is to propose an enlarged vision of the homogenization of equation (10.1) through the construction, for any $K \in \mathbb{N}$, of a homogenized equation of order $2K + 2$,

$$\begin{cases} \sum_{k=0}^{K+1} \epsilon^{2k-2} \mathbb{D}_K^{2k} \cdot \nabla^{2k} v_{\epsilon,K}^* = f \quad \text{in } D, \\ v_{\epsilon,K}^* \text{ is } D\text{-periodic}, \end{cases} \tag{10.6}$$

which yields an approximation of $u_\epsilon$ of order $O(\epsilon^{2K+4})$ in the $L^2(D_\epsilon)$ norm (for a fixed given scaling of the obstacles $\eta$). The function $v_{\epsilon,K}^*$ denotes the higher-order homogenized approximation of $u_\epsilon$, and $\mathbb{D}_K^{2k} \cdot \nabla^{2k}$ is a differential operator of order $2k$ with constant coefficients (the notation is defined in equation (10.17)). Equation (10.6) is a "corrected" version of the zeroth-order model (10.5) (for any $K$, it holds $\mathbb{D}_K^0 = M^0$), which yields a more accurate solution when $\epsilon$ is "not so small".

Our mathematical methodology is inspired from the works of Bakhvalov and Panasenko [15], Smyshlyaev and Cherednichenko [52], and Allaire et al. [12]; it starts with the identification of a "classical" two-scale ansatz

$$u_\epsilon(x) = \sum_{i=0}^{+\infty} \epsilon^{i+2} u_i(x, x/\epsilon), \quad x \in D_\epsilon, \tag{10.7}$$

expressed in terms of $Y$-periodic functions $u_i : D \times Y \to \mathbb{R}$ that do not depend on $\epsilon$. Our procedure involves then formal operations on related power series, which give rise to several families of tensors and homogenized equations for approximating the formal infinite-order homogenized average $u_\epsilon^*$:

$$u_\epsilon^*(x) := \sum_{i=0}^{+\infty} \epsilon^{i+2} \int_Y u_i(x, y) \mathrm{d}y, \quad x \in D. \tag{10.8}$$

In Proposition 10.3.5, we obtain that $u_\epsilon^*$ in equation (10.11) is the solution of a formal "infinite-order" homogenized equation,

$$\sum_{k=0}^{+\infty} \epsilon^{2k-2} M^{2k} \cdot \nabla^{2k} u_\epsilon^* = f, \tag{10.9}$$

where $M^0$ is the positive constant of equation (10.5), and $(M^k)_{k\geq 1}$ is a family of (constant) tensors or order $k$. From a computational point of view, we need a well-posed finite-order model. As it can be expected from other physical contexts [9, 12], the effective model obtained from a naive truncation of equation (10.9), say at order $2K$,

$$\sum_{k=0}^{K} \epsilon^{2k-2} M^{2k} \cdot \nabla^{2k} v_{\epsilon,K}^* = f \tag{10.10}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

is in general not well posed [1, 9]. Several techniques have been proposed in the literature to obtain well-posed homogenized models of finite order in the context of the conductivity or of the wave equation [1, 2, 10, 12, 13]. The derivation of the *well-posed* homogenized equation (10.6) relies on a minimization principle inspired by Smyshlyaev and Cherednichenko [52] and is marked by two surprising facts.

The first surprising result is the existence of a somewhat remarkable identity, which expresses the oscillating solution $u_\epsilon$ in terms of its nonoscillating average $u_\epsilon^*$:

$$u_\epsilon(x) = \sum_{k=0}^{+\infty} \epsilon^k N^k(x/\epsilon) \cdot \nabla^k u_\epsilon^*(x). \tag{10.11}$$

The functions $N^k$ are $P$-periodic corrector tensors of order $k$ (Definition 10.3.2) depending only on the shape of the obstacles $\eta T$ and that vanish on $\partial(\eta T)$. Furthermore, $N^0$ is of average $\int_Y N^0(y)\mathrm{d}y = 1$, and $N^k$ is of average $\int_Y N^k(y)\mathrm{d}y = 0$ for $k \geq 1$ (Proposition 10.3.8): $u_\epsilon^*$ is consistently the average of $u_\epsilon$ with respect to the fast variable $x/\epsilon$. Although the derivation of (10.7) is very standard in periodic homogenization [14, 18, 41, 50], the existence of such relation (10.11) (when compared to (10.7)) between the oscillating solution is less obvious; it has been noticed for the first time by Bakhvalov and Panasenko [15] for the conductivity equation and then in further homogenization contexts in [1, 2, 12, 23, 51, 52]. Following the denomination of [12], we call the ansatz (10.11) "criminal" because the function $u_\epsilon^*$ has the structure of a formal power series in $\epsilon$.

The second surprise lies in that our higher-order homogenized equation (10.6) is obtained by adding to (10.10) a *single* term $\epsilon^{2K}\mathbb{D}_K^{2K+2}\cdot\nabla^{2K+2}$; in other words, $\mathbb{D}_K^{2k} = M^{2k}$ for any $0 \leq 2k \leq 2K$. This fact, which does not seem to have been noticed in previous works, is quite surprising because following [23, 52], the derivation of (10.6) is based on a minimization principle for the truncation of (10.11) at order $K$,

$$W_{\epsilon,K}(v_{\epsilon,K}^*)(x) := \sum_{k=0}^{K} \epsilon^k N^k(x/\epsilon) \cdot \nabla^k v_{\epsilon,K}^*(x), \quad x \in D_\epsilon, \tag{10.12}$$

which is expected to yield an approximation of order $O(\epsilon^{K+3})$ only in the $L^2(D_\epsilon)$ norm ($u_\epsilon^*$ and $v_{\epsilon,K}^*$ are of order $O(\epsilon^2)$). This is the order of accuracy stated in our previous work [33] and similarly obtained in the conductivity case by [52], or for the Maxwell equations, by [23]; it is related to the observation that the first half of the coefficients of (10.6) and (10.9) coincide: $\mathbb{D}_K^{2k} = M^{2k}$ for any $0 \leq 2k \leq K$ (Proposition 10.4.4). In fact, it turns out that *all coefficients* $\mathbb{D}_K^{2k}$ *and* $M^{2k}$ *coincide* except the one of the leading order; $\mathbb{D}_K^{2K+2} \neq M^{2K+2}$. As a result, we are able to show in the present paper that the reconstructed function obtained by adding more correctors,

$$W_{\epsilon,2K+1}(v_{\epsilon,K}^*)(x) := \sum_{k=0}^{2K+1} \epsilon^k N^k(x/\epsilon) \cdot \nabla^k v_{\epsilon,K}^*(x), \quad x \in D_\epsilon, \tag{10.13}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

yields an approximation of $u_\epsilon$ of order $O(\epsilon^{2K+4})$ (Corollary 10.4.2):

$$\|u_\epsilon - W_{\epsilon,2K+1}(v^*_{\epsilon,K})\|_{L^2(D_\epsilon)} + \epsilon\|\nabla(u_\epsilon - W_{\epsilon,2K+1}(v^*_{\epsilon,K}))\|_{L^2(D_\epsilon)} \leq C_K(f)\epsilon^{2K+4}. \tag{10.14}$$

Finally, we obtain in Corollaries 10.5.1 and 10.5.2 (see also Remark 10.5.3) that our homogenized models have the potential to "unify" the different regimes of the literature, in the sense that equations (10.6) and (10.9) converge formally (coefficientwise) to either of the effective equations (10.3) and (10.4) when the scaling of the obstacle $\eta \to 0$ vanishes at rates respectively equivalent to or greater than the critical size $\eta_{\text{crit}} \sim \eta^{2/(d-2)}$.

Unfortunately, we do not obtain in this work that this convergence holds for all possible rates, because the estimates of Corollary 10.5.1 imply that higher-order coefficients $\epsilon^{2k-2}M^{2k}$ with $k > 2$ can blow up if the rate $\eta$ vanishes faster than the critical size (i. e., when $\eta = o(\epsilon^{2/(d-2)})$). However, the coefficientwise convergence holds for the homogenized equation (10.6) of order 2 (with $K = 0$). Although (i) the derivation of (10.6) has been performed by assuming $\eta$ constant and (ii) all our error bounds feature constants $C_K(f)$ which depend a priori on $\eta$, these results seem to indicate that (10.6) has the potential to yield valid homogenized approximations of (10.1) in any regime of size of holes if $K = 0$ (which was our initial goal) and for any size $\eta \geq \eta_{\text{crit}}$ if $K \geq 1$.

The exposure of our work outlines as follows. In Section 10.2, we introduce the notation conventions and provide a brief summary of our derivations.

In Section 10.3, we detail the procedure that allows us to construct the family of tensors $M^k$ and $N^k(y)$ arising in the formal infinite-order homogenized equation (10.9) and in the criminal ansatz (10.11). Additionally, we establish a number of algebraic properties satisfied by these tensors and provide an account of the simplifications that occur in case of symmetries of the obstacle with respect to the unit cell axes.

Section 10.4 is devoted to the construction of the finite-order homogenized equation (10.6) thanks to the method of Smyshlyaev and Cherednichenko. We prove the ellipticity of the model, and we establish that $\mathbb{D}_K^{2k} = M^{2k}$ for any $0 \leq 2k \leq 2K$ (and not only for the first half coefficients with $0 \leq 2k \leq K$ as observed in [33, 52]). The high-order homogenization process is then properly justified by establishing the error estimate (10.14).

Finally, in Section 10.5, we examine the asymptotic properties of the tensors $M^k$ in the low-volume fraction limit as $\eta \to 0$ (in space dimension $d \geq 3$). This allows us to retrieve formally the classical regimes and the arising of the celebrated "strange term" (see [25]) at the critical scaling $\eta \sim \epsilon^{2/(d-2)}$.

## 10.2 Notation and summary of the derivation

The full derivation of higher-order homogenized equations involves the construction of a number of families of tensors such as $\mathcal{X}^k$, $M^k$, $N^k$, $\mathbb{D}_K^{2k}$. For the convenience of the

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

reader, the notation conventions related to two-scale functions and tensor operations are summarized in Section 10.2.1. We then provide a short synthesis of our main results and of the key steps of our derivations in Section 10.2.2.

## 10.2.1 Notation conventions

Below and further on, we consider scalar functions such as

$$\begin{aligned} u: \quad & D \times P \to \mathbb{R} \\ & (x,y) \mapsto u(x,y) \end{aligned} \tag{10.15}$$

which are both $D$- and $P$-periodic with respect to, respectively, the first and second variables, and which vanish on the hole $D \times (\eta T)$. The arguments $x$ and $y$ of $u(x,y)$ are respectively called the "slow" and "fast" or "oscillating" variable. With a small abuse of notation, the partial derivative with respect to the variable $y_j$ (respectively, $x_j$) is simply written $\partial_j$ instead of $\partial_{y_i}$ (respectively, $\partial_{x_j}$) when the context is clear, i. e., when the function to which it is applied depends only on $y$ (respectively, only on $x$).

The star symbol $*$ is used to indicate that a quantity is "macroscopic" in the sense it does not depend on the fast variable $x/\epsilon$, e. g., $v^*_{\epsilon,K}$ in (10.6), $u^*_\epsilon$ in (10.8), or $J^*_K$ in (10.29). In the particular case where a two-variable quantity $u(x,y)$ is given such as (10.15), $u^*(x)$ always denotes the average of $y \mapsto u(x,y)$ with respect to the $y$ variable:

$$u^*(x) := \int_P u(x,y)\mathrm{d}y = \int_Y u(x,y)\mathrm{d}y, \quad x \in D,$$

where the last equality is a consequence of $u$ vanishing on $P \setminus Y = \overline{\eta T}$.

When a function $\mathcal{X} : P \to \mathbb{R}$ depends only on the $y$ variable, we find sometimes more convenient (especially, in Section 10.5) to write its cell average with the usual angle bracket symbols:

$$\langle \mathcal{X} \rangle := \int_P \mathcal{X}(y)\mathrm{d}y.$$

In all what follows and unless otherwise specified, the Einstein summation convention over repeated *subscript* indices is assumed (but never on *superscript* indices). Vectors $\boldsymbol{b} \in \mathbb{R}^d$ are written in bold face notation.

The notation conventions including those used related to tensor are summarized in the nomenclature below.

| | |
|---|---|
| $\boldsymbol{b}$ | Vector of $\mathbb{R}^d$. |
| $(b_j)_{1\le j\le d}$ | Coordinates of the vector $\boldsymbol{b}$. |
| $b^k$ | Tensor of order $k$ ($b^k_{i_1\dots i_k} \in \mathbb{R}$ for $1 \le i_1, \dots, i_k \le d$). |

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$b^p \otimes c^{k-p}$ Tensor product of tensors of order $p$ and $k-p$:

$$(b^p \otimes c^{k-p})_{i_1 \dots i_k} := b^p_{i_1 \dots i_p} c^{k-p}_{i_{p+1} \dots i_k}. \tag{10.16}$$

$b^k \cdot \nabla^k$ Differential operator of order $k$ associated with a tensor $b^k$:

$$b^k \cdot \nabla^k := b^k_{i_1 \dots i_k} \partial^k_{i_1 \dots i_k}, \tag{10.17}$$

with implicit summation over the repeated indices $i_1 \dots i_k$.

$(\boldsymbol{e}_j)_{1 \le j \le d}$ Vectors of the canonical basis of $\mathbb{R}^d$.

$e_j$ Tensor of order 1 whose entries are $(\delta_{i_1 j})_{1 \le i_1 \le d}$ (for any $1 \le j \le d$). Note: $e_j$ and $\boldsymbol{e}_j$ are the same mathematical object when identifying tensors of order 1 to vectors in $\mathbb{R}^d$.

$\delta_{ij}$ Kronecker symbol: $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$.

$I$ Identity tensor of order 2:

$$I_{i_1 i_2} = \delta_{i_1 i_2}.$$

Note that the identity tensor is another notation for the Kronecker tensor and it holds $I = e_j \otimes e_j$.

$J^{2k}$ Tensor of order $2k$ defined by:

$$J^{2k} := \overbrace{I \otimes I \otimes \cdots \otimes I}^{k \text{ times}}. \tag{10.18}$$

$\mathbb{B}^{k,l}_K$ Tensor of order $l+m$ associated with the quadratic form

$$\mathbb{B}^{l,m}_K \nabla^l v \nabla^m w := \mathbb{B}^{l,m}_{l,i_1 \dots i_l, j_1 \dots j_m} \partial^l_{i_1 \dots i_l} v \, \partial^m_{j_1 \dots j_m} w, \tag{10.19}$$

for any smooth scalar fields $v, w \in \mathcal{C}^\infty(D)$.

With a small abuse of notation, we consider zeroth-order tensors $b^0$ to be constants (i. e., $b^0 \in \mathbb{R}$), and we still denote by $b^0 \otimes c^k := b^0 c^k$ the tensor product with a $k$th-order tensor $c^k$.

In all what follows, a $k$th order tensor $b^k$ truly makes sense when contracted with $k$ partial derivatives, as in (10.17). Therefore all the tensors considered throughout this work are identified to their symmetrization:

$$b^k_{i_1 \dots i_k} \equiv \frac{1}{k!} \sum_{\sigma \in \mathfrak{S}_k} b_{i_{\sigma(1)} \dots i_{\sigma(k)}},$$

where $\mathfrak{S}_k$ is the permutation group of order $k$. Consequently, the order in which the indices $i_1, \dots, i_k$ are written in $b^k_{i_1 \dots i_k}$ does not matter, and the tensor product $\otimes$ is com-

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

mutative under this identification:

$$b^k \otimes c^{k-p} = c^{k-p} \otimes b^k. \tag{10.20}$$

Finally, $C$, $C_K$, or $C_K(f)$ denote universal constants that do not depend on $\epsilon$ but whose values may change from line to line (and which depend a priori on the shape of the hole $\eta T$).

### 10.2.2 Summary of the derivation

One of the main results of this paper is the derivation of the higher-order homogenized equation (10.6) for any desired order $K \in \mathbb{N}$ and the justification of the procedure by establishing the error estimate (10.14). Before summarizing the most essential steps of our analysis, let us recall that here and in all what follows, equalities involving infinite power series such as (10.11) are formal and without a precise meaning of convergence. Our derivation outlines as follows:

1. Following the classical literature [15, 19, 40, 41], we introduce a family of $k$th-order tensors $(\mathcal{X}^k(y))_{k\in\mathbb{N}}$ obtained as the solutions of cell problems (see Definition 10.3.1 and Proposition 10.3.1), which allows us to identify the functions $u_i(x,y)$ arising in (10.11) and to rewrite the traditional ansatz more explicitly:

$$u_\epsilon(x) = \sum_{i=0}^{+\infty} \epsilon^{i+2} \mathcal{X}^i(x/\epsilon) \cdot \nabla^i f(x), \quad x \in D_\epsilon. \tag{10.21}$$

   Introducing the averaged tensors of order $i$, $\mathcal{X}^{i*} := \int_Y \mathcal{X}^i(y)\mathrm{d}y$, the formal average (10.8) reads

$$u_\epsilon^*(x) = \sum_{i=0}^{+\infty} \epsilon^{i+2} \mathcal{X}^{i*} \cdot \nabla^i f(x). \tag{10.22}$$

2. We construct (in Proposition 10.3.5) constant tensors $M^i$ by inversion of the formal equality

$$\left(\sum_{i=0}^{+\infty} \epsilon^{i-2} M^i \cdot \nabla^i\right)\left(\sum_{i=0}^{+\infty} \epsilon^{i+2} \mathcal{X}^{i*} \cdot \nabla^i\right) = I,$$

   which yields, after left multiplication of (10.22) by $\sum_{i=0}^{+\infty} \epsilon^{i-2} M^i \cdot \nabla^i$, the "infinite-order homogenized equation" for $u_\epsilon^*(x)$:

$$\sum_{i=0}^{+\infty} \epsilon^{i-2} M^i \cdot \nabla^i u_\epsilon^*(x) = f(x). \tag{10.23}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Note that equation (10.23) is exactly equation (10.9) because all tensors $\mathcal{X}^{2k+1*}$ and $M^{2k+1}$ of odd order vanish (Proposition 10.3.3 and Corollary 10.3.1).

3. We substitute the expression of $f(x)$ given by equation (10.23) into the ansatz (10.21) so as to recognize a formal series product:

$$u_\epsilon(x) = \left(\sum_{i=0}^{+\infty} \epsilon^i \mathcal{X}^i(x/\epsilon)\cdot\nabla^i\right)\left(\sum_{i=0}^{+\infty} \epsilon^i M^i\cdot\nabla^i\right)u_\epsilon^*(x). \tag{10.24}$$

Introducing a new family of tensors $N^k(y)$ defined by the corresponding Cauchy product,

$$N^k(y) := \sum_{p=0}^{k} \mathcal{X}^p(y)\otimes M^{k-p}, \quad y\in Y,$$

we obtain the "criminal" ansatz (10.11):

$$u_\epsilon(x) = \sum_{k=0}^{+\infty} \epsilon^k N^k(x/\epsilon)\cdot\nabla^k u_\epsilon^*(x). \tag{10.25}$$

4. We now seek to construct well-posed effective models of finite order. Inspired by [23, 52], we consider truncated versions of functions of the form (10.25): for any $v\in H^{K+1}(D)$, we define

$$W_{\epsilon,K}(v)(x) := \sum_{k=0}^{K} \epsilon^k N^k(x/\epsilon)\cdot\nabla^k v(x), \quad x\in D. \tag{10.26}$$

where the tensors $N^k$ are extended by zero in $P\setminus Y$ for this expression to make sense in $D\backslash D_\epsilon$. Recalling that $u_\epsilon$ (identified with its extension by 0 in $D\backslash D_\epsilon$) is the solution to the energy minimization problem

$$\begin{aligned} &\min_{u\in H^1(D)} \quad J(u,f) := \int_D \left(\frac{1}{2}|\nabla u|^2 - fu\right)\mathrm{d}x \\ &\text{s. t.} \quad \begin{cases} u = 0 \text{ in } \omega_\epsilon, \\ u \text{ is } D\text{-periodic}, \end{cases} \end{aligned} \tag{10.27}$$

we formulate an analogous minimization problem for $v\in H^{K+1}(D)$ by restriction of (10.27) to the smaller space of functions $W_{\epsilon,K}(v)\in H^1(D)$:

$$\begin{aligned} &\min_{v\in H^{K+1}(D)} \quad J(W_{\epsilon,K}(v),f) = \int_D \left(\frac{1}{2}\left|\nabla W_{\epsilon,K}(v)\right|^2 - f(x)W_{\epsilon,K}(v)(x)\right)\mathrm{d}x \\ &\text{s. t.} \quad v \text{ is } D\text{-periodic}. \end{aligned} \tag{10.28}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Averaging over the fast variable $x/\epsilon$ (by using Lemma 10.4.3), we obtain a new minimization problem involving an approximate energy $J_K^*$ (Definition 10.4.2), which does not depend on $x/\epsilon$,

$$\begin{aligned} \min_{v \in H^{K+1}(D)} \quad & J_K^*(v, f, \epsilon) \\ \text{s. t.} \quad & v \text{ is } D\text{-periodic.} \end{aligned} \tag{10.29}$$

Its Euler–Lagrange equation (see Definition 10.4.3) finally defines our well-posed homogenized equation (10.6) and, in particular, the family of tensors $(\mathbb{D}_K^{2k})_{0 \le 2k \le 2K++22}$.

5. In view of (10.26), this procedure is expected to yield by construction an approximation $W_{\epsilon,K}(v_{\epsilon,K}^*)$ of $u_\epsilon$ with an error $O(\epsilon^{K+3})$ (because $v_{\epsilon,K}^*$ is of order $O(\epsilon^2)$, see Lemma 10.5.2). Surprisingly, we verify that all the tensors $\mathbb{D}_K^{2k}$ and $M^{2k}$ coincide for $0 \le 2k \le 2K$ (Proposition 10.4.4). This allows us to obtain in Corollary 10.4.2 the error estimate (10.14), which states that $v_{\epsilon,K}^*$ and reconstruction (10.13) yield a *much better* approximation than expected, namely of order $O(\epsilon^{2K+4})$ instead of $O(\epsilon^{K+3})$.

The most essential point of our methodology is the derivation of the non-classical ansatz equation (10.25) of step (3). Let us stress that in the available works of the literature concerned with high-order homogenization of scalar conductivity equations and its variants [12, 15, 52], the criminal ansatz (analogous to (10.25)) is readily obtained from the classical one (analogous to (10.11)) because the tensors $N^k$ and $\mathcal{X}^k$ coincide in these contexts (check, for instance, [9, 15]). Our case is very different because the heterogeneity comes from the Dirichlet boundary condition on the holes $\omega_\epsilon$.

## 10.3 Derivation of the infinite-order homogenized equation and of the criminal ansatz

In this section we present the steps (1) to (3) of Section 10.2.2 in detail. We start in Subsection 10.3.1 by reviewing the definition of the family of cell tensors $(\mathcal{X}^k(y))_{k \in \mathbb{N}}$, which allows us to identify the functions $u_i(x, y)$ involved in the "usual" two-scale ansatz (10.8) and to obtain an error estimate for its truncation in Proposition 10.3.2. This part is not new; it is a review of classical results available in a number of works; see, e. g., [19, 40, 41]. We then establish in Section 10.3.2 several properties of the tensor $\mathcal{X}^k$ that are less found in the literature, the most important being that the averages of all odd order tensors vanish: $\mathcal{X}^{2p+1*} = 0$ for any $p \in \mathbb{N}$.

The next Section 10.3.3 focuses on the definition and properties of the family of tensors $M^k$ and $N^k(y)$, which allow us to infer the infinite-order homogenized equation (10.9) and the criminal ansatz (10.11).

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Finally, we investigate in Section 10.3.4 how symmetries of the obstacle $\eta T$ with respect to the axes of the unit cell $P$ reflect into a decrease of the number of independent components of the homogenized tensors $\mathcal{X}^{k*}$ and $M^k$.

### 10.3.1 Traditional ansatz: definition of the cell tensors $\mathcal{X}^k$

Classically [40, 41, 50], the first step of our analysis is to insert formally the two-scale expansion (10.11) into the Poisson system (10.1). Because it will help highlight the occurrence of Cauchy products, we also assume (for the purpose of the derivation only) that the right-hand side $f \in \mathcal{C}^\infty(D)$ depends on $\epsilon$ and admits the following formal expansion:

$$f(x) = \sum_{i=0}^{+\infty} \epsilon^i f_i(x), \quad x \in D.$$

Evaluating the Laplace operator against (10.11), we formally obtain

$$-\Delta u_\epsilon = \sum_{i=-2}^{+\infty} \epsilon^{i+2}(-\Delta_{yy} u_{i+2} - \Delta_{xy} u_{i+1} - \Delta_{xx} u_i),$$

where we use the convention $u_{-2}(x,y) = u_{-1}(x,y) = 0$, and where $-\Delta_{yy}$, $-\Delta_{xy}$, and $-\Delta_{xx}$ are the operators

$$-\Delta_{xx} := -\operatorname{div}_x(\nabla_x \cdot), \quad -\Delta_{xy} := -\operatorname{div}_x(\nabla_y \cdot) - \operatorname{div}_y(\nabla_x \cdot), \quad -\Delta_{yy} := -\operatorname{div}_y(\nabla_y \cdot).$$

Then identifying all powers in $\epsilon$ yields the traditional cascade of equations (obtained, e. g., in [41]):

$$\begin{cases} -\Delta_{yy} u_{i+2} = f_{i+2} + \Delta_{xy} u_{i+1} + \Delta_{xx} u_i \quad \text{for all } i \geq -2, \\ u_{-2}(x,y) = u_{-1}(x,y) = 0. \end{cases} \tag{10.30}$$

This system of equations is solved by introducing an appropriate family of cell tensors [40, 41].

**Definition 10.3.1.** We define the family of tensors $(\mathcal{X}^{(}y))_{k\in\mathbb{N}}$ of order $k$ by recurrence as follows:

$$\begin{cases} -\Delta_{yy} \mathcal{X}^0 = 1 \quad \text{in } Y, \\ -\Delta_{yy} \mathcal{X}^1 = 2\partial_j \mathcal{X}^0 \otimes e_j \quad \text{in } Y, \\ -\Delta_{yy} \mathcal{X}^{k+2} = 2\partial_j \mathcal{X}^{k+1} \otimes e_j + \mathcal{X}^k \otimes I \quad \text{in } Y \text{ for all } k \geq 0, \\ \mathcal{X}^k = 0 \quad \text{on } \partial(\eta T), \\ \mathcal{X}^k \text{ is } P\text{-periodic}. \end{cases} \tag{10.31}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

The tensors $\mathcal{X}^k$ are extended by 0 inside the hole $\eta T$ in the whole unit cell $P$, namely, $\mathcal{X}^k(y) = 0$ for $y \in \eta T$.

**Remark 10.3.1.** In view of (10.16), the third line of (10.31) is a short-hand notation for

$$-\Delta \mathcal{X}^{k+2}_{i_1 \dots i_{k+2}} = 2\partial_{i_{k+2}} \mathcal{X}^{k+1}_{i_1 \dots i_{k+1}} + \mathcal{X}^{k}_{i_1 \dots i_k} \delta_{i_{k+1} i_{k+2}} \quad \text{for all } k \geq 0.$$

**Proposition 10.3.1.** *The solutions $(u_i(x,y))_{i\geq 0}$ to the cascade of equations* (10.30) *are given by*

$$\forall i \geq 0, \quad u_i(x,y) = \sum_{k=0}^{i} \mathcal{X}^k(y) \cdot \nabla^k f_{i-k}(x), \quad x \in D,\ y \in Y. \tag{10.32}$$

*Recognizing a Cauchy product, the ansatz* (10.11) *can be formally written as the following infinite power series product:*

$$u_\epsilon(x) = \sum_{i=0}^{+\infty} \epsilon^{i+2} \mathcal{X}^i(x/\epsilon) \cdot \nabla^i f(x) = \epsilon^2 \left( \sum_{i=0}^{+\infty} \epsilon^i \mathcal{X}^i(x/\epsilon) \cdot \nabla^i \right) \left( \sum_{i=0}^{+\infty} \epsilon^i f_i(x) \right). \tag{10.33}$$

*Proof.* See [33, 40, 41]. □

We complete our review by stating a classical error estimate result, which justifies in some sense the formal power series expansion (10.33).

**Proposition 10.3.2.** *Denote by $u_{\epsilon,K}$ the truncated ansatz of* (10.33) *at order $K \in \mathbb{N}$:*

$$u_{\epsilon,K}(x) := \sum_{i=0}^{K} \epsilon^{i+2} \mathcal{X}^i(x/\epsilon) \cdot \nabla^i f(x), \quad x \in D_\epsilon.$$

*Then assuming that $f \in \mathcal{C}^\infty(D)$ is $D$-periodic, we have the following error bound:*

$$\|u_\epsilon - u_{\epsilon,K}\|_{L^2(D_\epsilon)} + \epsilon \left\| \nabla(u_\epsilon - u_{\epsilon,K}) \right\|_{L^2(D_\epsilon)} \leq C_K \epsilon^{K+3} \|f\|_{H^{K+2}(D)}$$

*for a constant $C_K$ independent of $f$ and $\epsilon$ (but depending on $K$).*

*Proof.* See [33, 40, 41]. □

## 10.3.2 Properties of the tensors $\mathcal{X}^k$: odd-order tensors $\mathcal{X}^{2p+1*}$ are zero

Following our conventions of Section 10.2.1, the average of the functions $u_i$ and $\mathcal{X}^i$ with respect to the $y$ variable are respectively denoted:

$$u_i^*(x) := \int_Y u_i(x,y) \mathrm{d}y, \quad x \in D, \tag{10.34}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$$\mathcal{X}^{i*} := \int_Y \mathcal{X}^i(y)\mathrm{d}y. \tag{10.35}$$

In the next proposition, we show that $\mathcal{X}^{2p+1*} = 0$ are of zero average for any $p \in \mathbb{N}$ and that $\mathcal{X}^{2p*}$ depends only on the lower-order tensors $\mathcal{X}^p$ and $\mathcal{X}^{p-1}$. Similar formulas have been obtained for the wave equation in heterogeneous media; see, e. g., Theorem 3.5 in [2] and also [1, 48].

**Proposition 10.3.3.** *For any $0 \le p \le k$, we have the following identity for the tensor $\mathcal{X}^{k*}$:*

$$\mathcal{X}^{k*} = \int_Y \mathcal{X}^k \mathrm{d}y = (-1)^p \int_Y (\mathcal{X}^{k-p} \otimes (-\Delta_{yy}\mathcal{X}^p) - \mathcal{X}^{k-p-1} \otimes \mathcal{X}^{p-1} \otimes I)\mathrm{d}y \tag{10.36}$$

*with the convention that $\mathcal{X}^{-1} = 0$. In particular, for any $p \in \mathbb{N}$:*

- $\mathcal{X}^{2p+1*} = 0$,
- *$\mathcal{X}^{2p*}$ depends only on the tensors $\mathcal{X}^p$ and $\mathcal{X}^{p-1}$:*

$$\mathcal{X}^{2p*} = (-1)^p \int_Y (\partial_j \mathcal{X}^p \otimes \partial_j \mathcal{X}^p - \mathcal{X}^{p-1} \otimes \mathcal{X}^{p-1} \otimes I)\mathrm{d}y. \tag{10.37}$$

*Proof.* We prove the result by induction. Formula (10.36) holds for $p = 0$ by using the convention $\mathcal{X}^{-1} = 0$ and $-\Delta_{yy}\mathcal{X}^0 = 1$. Assuming now the result to be true for $0 \le p < k$, we perform the following integration by parts, where we use the boundary conditions satisfied by the tensors $\mathcal{X}^k$ and the commutativity property (10.20) of the tensor product:

$$\begin{aligned}
\mathcal{X}^{k*} &= (-1)^p \int_Y (-\Delta_{yy}\mathcal{X}^{k-p} \otimes \mathcal{X}^p - \mathcal{X}^{k-p-1} \otimes \mathcal{X}^{p-1} \otimes I)\mathrm{d}y \\
&= (-1)^p \int_Y ((2\partial_j \mathcal{X}^{k-p-1} \otimes e_j + \mathcal{X}^{k-p-2} \otimes I) \otimes \mathcal{X}^p - \mathcal{X}^{k-p-1} \otimes \mathcal{X}^{p-1} \otimes I)\mathrm{d}y \\
&= (-1)^p \int_Y (-2\partial_j \mathcal{X}^p \otimes e_j - \mathcal{X}^{p-1} \otimes I) \otimes \mathcal{X}^{k-p-1} + \mathcal{X}^{k-p-2} \otimes \mathcal{X}^p \otimes I)\mathrm{d}y \\
&= (-1)^{p+1} \int_Y ((-\Delta_{yy}\mathcal{X}^{p+1}) \otimes \mathcal{X}^{k-p-1} - \mathcal{X}^{k-p-2} \otimes \mathcal{X}^p \otimes I)\mathrm{d}y.
\end{aligned}$$

Hence the formula is proved for order $p + 1$.

Now the formula of order $p = k$ reads

$$\mathcal{X}^{k*} = (-1)^k \int_Y \mathcal{X}^0(-\Delta_{yy}\mathcal{X}^k)\mathrm{d}y = (-1)^k \int_Y \mathcal{X}^k(-\Delta_{yy}\mathcal{X}^0)\mathrm{d}y = (-1)^k \mathcal{X}^{k*},$$

which implies that $\mathcal{X}^{k*} = 0$ if $k$ is odd. Formula (10.37) follows easily from (10.36) with $k = 2p$. □

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

For completeness, we provide a minor result, which implies that there is no order $k$ (even odd) such that $\mathcal{X}^k(y)$ is identically equal to zero. However, let us remark that some components $\mathcal{X}^k_{i_1\dots i_k}(y)$ may vanish for some set of indices $i_1,\dots,i_k$, e. g., in case of invariance of the obstacle $\eta T$ along the cell axes.

**Proposition 10.3.4.** *We have the following identity:*

$$-\Delta_{yy}(\partial^k_{i_1\dots i_k}\mathcal{X}^k_{i_1\dots i_k}) = (-1)^k(k+1),$$

*where we recall the implicit summation convention over the repeated indices $i_1\dots i_k$.*

*Proof.* The results clearly holds for $k=0$. For $k=1$, we have

$$-\Delta_{yy}\partial_i\mathcal{X}^1_i = \partial_i(2\partial_i\mathcal{X}^0) = 2\Delta\mathcal{X}^0 = -2.$$

Assuming that the result holds till rank $k-1$, the formula still holds at rank $k\geq 2$ because

$$\begin{aligned}
-\Delta_{yy}\partial^k_{i_1\dots i_k}\mathcal{X}^k_{i_1\dots i_k} &= \partial^k_{i_1\dots i_k}(2\partial_{i_k}\mathcal{X}^{k-1}_{i_1\dots i_{k-1}} + \mathcal{X}^{k-2}_{i_1\dots i_{k-2}}\delta_{i_{k-1}i_k})\\
&= 2\Delta_{yy}(\partial^{k-1}_{i_1\dots i_{k-1}}\mathcal{X}^{k-1}_{i_1\dots i_{k-1}}) + \Delta_{yy}(\partial^{k-2}_{i_1\dots i_{k-2}}\mathcal{X}^{k-2}_{i_1\dots i_{k-2}})\\
&= -2(-1)^{k-1}k - (-1)^{k-2}(k-1)\\
&= (-1)^k(k+1).
\end{aligned}$$

□

### 10.3.3 Infinite-order homogenized equation and criminal ansatz: tensors $M^k$ and $N^k$

This part outlines steps (2) and (3) of the procedure outlined in Section 10.2.2. Let us recall that the first tensor $\mathcal{X}^{0*}$ is a strictly positive number, since (10.37) implies $\mathcal{X}^{0*} = \int_Y |\nabla\mathcal{X}^0|^2\mathrm{d}y > 0$.

**Proposition 10.3.5.** *Let $(M^k)_{i\in\mathbb{N}}$ be the family of kth-order tensors defined by induction as follows:*

$$\begin{cases} M^0 = (\mathcal{X}^{0*})^{-1},\\ M^k = -(\mathcal{X}^{0*})^{-1}\sum_{p=0}^{k-1}\mathcal{X}^{k-p*}\otimes M^p. \end{cases} \tag{10.38}$$

*Then, given definitions* (10.30) *and* (10.34) *of $u_i^*$,*

$$\forall i\in\mathbb{N},\quad f_i(x) = \sum_{k=0}^{i} M^k\cdot\nabla^k u^*_{i-k}(x). \tag{10.39}$$

*Recognizing a Cauchy product, formula* (10.39) *can be rewritten formally in terms of the following "infinite-order" homogenized equation for the "infinite-order" homogenized*

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*average $u_\epsilon^*$ of* (10.22)*:*

$$\sum_{i=0}^{+\infty} \epsilon^{i-2} M^i \cdot \nabla^i u_\epsilon^* = f. \tag{10.40}$$

*Proof.* We proceed by induction. The case $i = 0$ results from the identity $u_0^*(x) = \mathcal{X}^{0*} f_0(x)$, which yields $f_0(x) = (\mathcal{X}^{0*})^{-1} u_0^*(x)$. Then, assuming that (10.39) holds till rank $i-1$ with $i \geq 1$, we average (10.32) with respect to the $y$ variable to obtain

$$u_i^* = \sum_{p=0}^{i} \mathcal{X}^{p*} \cdot \nabla^p f_{i-p} = \mathcal{X}^{0*} f_i + \sum_{p=1}^{i} \mathcal{X}^{p*} \cdot \nabla^p f_{i-p}.$$

By using formula (10.39) at ranks $i-p$ with $1 \leq p \leq i$, we obtain the following expression for $f_i$:

$$\begin{aligned}
f_i &= (\mathcal{X}^{0*})^{-1} \left( u_i^* - \sum_{p=1}^{i} \sum_{q=0}^{i-p} (\mathcal{X}^{p*} \otimes M^q) \cdot \nabla^{p+q} u_{i-p-q}^* \right) \\
&= (\mathcal{X}^{0*})^{-1} \left( u_i^* - \sum_{p=1}^{i} \sum_{k=p}^{i} (\mathcal{X}^{p*} \otimes M^{k-p}) \cdot \nabla^k u_{i-k}^* \right) \quad (\text{by } k \leftarrow p+q) \\
&= (\mathcal{X}^{0*})^{-1} \left( u_i^* - \sum_{k=1}^{i} \sum_{p=1}^{k} (\mathcal{X}^{p*} \otimes M^{k-p}) \cdot \nabla^k u_{i-k}^* \right) \quad (\text{inversion of summation}) \\
&= (\mathcal{X}^{0*})^{-1} \left( u_i^* - \sum_{k=1}^{i} \left( \sum_{p=0}^{k-1} \mathcal{X}^{k-p*} \otimes M^p \right) \cdot \nabla^k u_{i-k}^* \right) \quad (\text{by } p \leftrightarrow k-p) \\
&= M^0 u_i^* + \sum_{k=1}^{i} M^k \cdot \nabla^k u_{i-k}^*,
\end{aligned}$$

which yields the result at rank $i$. □

**Corollary 10.3.1.** *$M^k = 0$ for any odd value of $k$.*

*Proof.* If $k$ is odd, then $k-p$ and $p$ have distinct parities in (10.38). Therefore the result follows by induction and by using $\mathcal{X}^{k-p*} = 0$ for even values of $p$ (Proposition 10.3.3). □

It is possible to write a more explicit formula for the tensors $M^k$.

**Proposition 10.3.6.** *The tensors $M^k$ are explicitly given by $M^0 = (\mathcal{X}^{0*})^{-1}$ and*

$$\forall k \geq 1, \quad M^k = \sum_{p=1}^{k} \frac{(-1)^p}{(\mathcal{X}^{0*})^{p+1}} \sum_{\substack{i_1+\cdots+i_p=k \\ 1 \leq i_1 \ldots i_p \leq k}} \mathcal{X}^{i_1 *} \otimes \cdots \otimes \mathcal{X}^{i_p *}. \tag{10.41}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*Proof.* For $k = 1$, the result is true because

$$M^1 = -\left(\mathcal{X}^{0*}\right)^{-1} M^0 \mathcal{X}^{1*} = -\left(\mathcal{X}^{0*}\right)^{-2} \mathcal{X}^{1*},$$

which is exactly formula (10.41). Assuming that (10.41) holds till rank $k \geq 1$, we compute

$$\begin{aligned}
M^{k+1} &= -\left(\mathcal{X}^{0*}\right)^{-1} \sum_{p=0}^{k} \mathcal{X}^{k+1-p*} \otimes M^p \\
&= -\left(\mathcal{X}^{0*}\right)^{-1} M^0 \mathcal{X}^{k+1*} \\
&\quad - \left(\mathcal{X}^{0*}\right)^{-1} \sum_{p=1}^{k} \sum_{q=1}^{p} \frac{(-1)^q}{(\mathcal{X}^{0*})^{q+1}} \mathcal{X}^{k+1-p*} \otimes \sum_{\substack{i_1+\cdots+i_q=p \\ 1\leq i_1 \ldots i_q \leq p}} \mathcal{X}^{i_1*} \otimes \cdots \otimes \mathcal{X}^{i_q*} \\
&= -\left(\mathcal{X}^{0*}\right)^{-2} \mathcal{X}^{k+1*} \\
&\quad - \left(\mathcal{X}^{0*}\right)^{-1} \sum_{q=1}^{k} \frac{(-1)^q}{(\mathcal{X}^{0*})^{q+1}} \sum_{p=q}^{k} \sum_{\substack{i_1+\cdots+i_q=p \\ 1\leq i_1 \ldots i_q \leq p}} \mathcal{X}^{k+1-p*} \otimes \mathcal{X}^{i_1*} \otimes \cdots \otimes \mathcal{X}^{i_q*} \\
&= -\left(\mathcal{X}^{0*}\right)^{-2} \mathcal{X}^{k+1*} \\
&\quad - \left(\mathcal{X}^{0*}\right)^{-1} \sum_{q=1}^{k} \frac{(-1)^q}{(\mathcal{X}^{0*})^{q+1}} \sum_{\substack{i_1+\cdots+i_{q+1}=k+1 \\ 1\leq i_1 \ldots i_{q+1} \leq k+1}} \mathcal{X}^{i_{q+1}*} \otimes \mathcal{X}^{i_1*} \otimes \cdots \otimes \mathcal{X}^{i_q*} \\
&= -\left(\mathcal{X}^{0*}\right)^{-2} \mathcal{X}^{k+1*} + \sum_{q=2}^{k+1} \frac{(-1)^q}{(\mathcal{X}^{0*})^{q+1}} \sum_{\substack{i_1+\cdots+i_q=k+1 \\ 1\leq i_1 \ldots i_q \leq k+1}} \mathcal{X}^{i_1*} \otimes \cdots \otimes \mathcal{X}^{i_q*},
\end{aligned}$$

from which the result follows. □

**Remark 10.3.2.** This result essentially states that $\sum_{k=0}^{+\infty} \epsilon^k M^k \cdot \nabla^k$ is the formal series expansion of

$$\left( \sum_{k=0}^{+\infty} \epsilon^k \mathcal{X}^k \cdot \nabla^k \right)^{-1}.$$

Indeed, it is elementary to show the following identity for the inverse of a power series $\sum_{k=0}^{+\infty} a_k z^k$ with $(a_k) \in \mathbb{C}^{\mathbb{N}}$, $z \in \mathbb{C}$, and radius of convergence $R > 0$:

$$\left( \sum_{k=0}^{+\infty} a_k z^k \right)^{-1} = a_0^{-1} + \sum_{k=1}^{+\infty} \left( \sum_{p=1}^{k} \frac{(-1)^p}{a_0^{p+1}} \sum_{\substack{i_1+\cdots+i_p=k \\ 1\leq i_1 \ldots i_p \leq k}} a_{i_1} a_{i_2} \ldots a_{i_p} \right) z^k. \tag{10.42}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

We now turn on the derivation of the "criminal" ansatz (10.25). Being guided by (10.24), this ansatz is obtained by writing the oscillatory part $u_i(x,y)$ in terms of the nonoscillatory part $u_i^*(x)$:

**Proposition 10.3.7.** *Given the previous definitions* (10.30) *and* (10.34) *of respectively* $u_i$ *and* $u_i^*$*, we have the following identity:*

$$\forall i \geq 0, \quad u_i(x,y) = \sum_{k=0}^{i} \left( \sum_{p=0}^{k} M^p \otimes \mathcal{X}^{k-p}(y) \right) \cdot \nabla^k u_{i-k}^*(x). \tag{10.43}$$

*Proof.* Substituting (10.39) into (10.32) yields

$$\begin{aligned} u_i(x,y) &= \sum_{p=0}^{i} \sum_{q=0}^{i-p} (\mathcal{X}^p(y) \otimes M^q) \cdot \nabla^{p+q} u_{i-p-q}^*(x) \\ &= \sum_{p=0}^{i} \sum_{k=p}^{i} (\mathcal{X}^p(y) \otimes M^{p-k}) \cdot \nabla^k u_{i-k}^*(x) \quad \text{(change of indices } k = p+q\text{)} \\ &= \sum_{k=0}^{i} \sum_{p=0}^{k} (\mathcal{X}^p(y) \otimes M^{p-k}) \cdot \nabla^k u_{i-k}^*(x) \quad \text{(interversion of summation).} \end{aligned}$$

The result follows by performing the last change of indices $p \leftrightarrow k-p$. □

This result motivates the definition of the tensors $N^k$ of equation (10.25):

**Definition 10.3.2.** For any $k \geq 0$, we denote by $N^k(y)$ the $k$th-order tensor

$$N^k(y) := \sum_{p=0}^{k} M^p \otimes \mathcal{X}^{k-p}(y), \quad y \in Y. \tag{10.44}$$

Recognizing a Cauchy product, identity (10.43) rewrites, as expected, as the "criminal" ansatz (10.25), which expresses the oscillating solution $u_\epsilon$ in terms of its formal homogenized averaged $u_\epsilon^*$ (defined in (10.22)):

$$u_\epsilon(x) = \sum_{k=0}^{+\infty} \epsilon^k N^k(x/\epsilon) \cdot \nabla^k u_\epsilon^*(x), \quad x \in D_\epsilon.$$

The last proposition of this section gathers several important properties for the tensors $N^k$ that are dual to those of the tensors $\mathcal{X}^k$ stated in Sections 10.3.1 and 10.3.2.

**Proposition 10.3.8.** *The tensor* $N^k(y)$ *satisfies:*

1. $\int_Y N^0(y)\mathrm{d}y = 1$ *and* $\int_Y N^k(y)\mathrm{d}y = 0$ *for any* $k \geq 1$.
2. *For any* $k \geq 0$*, assuming the convention* $N^{-1} = N^{-2} = 0$*,*

$$-\Delta_{yy} N^{k+2} = 2\partial_j N^{k+1} \otimes e_j + N^k \otimes I + M^{k+2}. \tag{10.45}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

3. *For any* $k \geq 0$,

$$-\Delta_{yy}(\partial^k_{i_1\dots i_k} N^k_{i_1\dots i_k}) = (-1)^k(k+1)M^0. \tag{10.46}$$

4. *For any* $k \geq 1$ *and* $1 \leq p \leq k-1$,

$$M^k = (-1)^{p+1}\int_Y (N^{k-p} \otimes (-\Delta_{yy}N^p) - N^{k-p-1} \otimes N^{p-1} \otimes I)\mathrm{d}y. \tag{10.47}$$

*In particular,* $M^{2p}$ *depends only on the tensors* $N^p$ *and* $N^{p-1}$*, which depend themselves only on the first* $p+1$ *tensors* $\mathcal{X}^0, \dots, \mathcal{X}^p$.

*Proof.*

1. For $k = 0$, we have $\int_Y N^0(y)\mathrm{d}y = M^0\mathcal{X}^{0*} = 1$. Furthermore, definition (10.38) of the tensor $M^k$ can be rewritten as

$$\forall k \geq 1, \quad \int_Y N^k(y)\mathrm{d}y = \sum_{p=0}^{k} \mathcal{X}^{k-p*} \otimes M^p = 0.$$

2. The cases $k = 0$ and $k = 1$ are easily verified. The case $k \geq 2$ is obtained by writing

$$\begin{aligned}
-\Delta_{yy}N^k &= \sum_{p=0}^{k} M^{k-p} \otimes (-\Delta_{yy}\mathcal{X}^p) \\
&= \sum_{p=2}^{k} M^{k-p} \otimes (2\partial_j\mathcal{X}^{p-1} \otimes e_j + \mathcal{X}^{p-2} \otimes I) + M^{k-1} \otimes (2\partial_j\mathcal{X}^0 \otimes e_j) + M^k \\
&= 2\partial_j\left(\sum_{p=1}^{k} M^{k-p} \otimes \mathcal{X}^{p-1}\right) \otimes e_j + \left(\sum_{p=0}^{k-2} M^p \otimes \mathcal{X}^{k-p-2}\right) \otimes I + M^k,
\end{aligned}$$

from which the result follows.

3. The proof of (10.46) is identical to that of Proposition 10.3.4.

4. We start by proving the result for $p = 1$ with $k > 1$: using item 1, we write

$$\begin{aligned}
M^k &= \int_Y N^0 \otimes M^k\mathrm{d}y = \int_Y N^0 \otimes (-\Delta N^k - 2\partial_j N^{k-1} \otimes e_j - N^{k-2} \otimes I)\mathrm{d}y \\
&= \int_Y M^0 \otimes N^k\mathrm{d}y + \int_Y (2\partial_j N^0 \otimes e_j \otimes N^{k-1} - N^0 \otimes N^{k-2} \otimes I)\mathrm{d}y \\
&= \int_Y ((-\Delta_{yy}N^1) \otimes N^{k-1} - N^0 \otimes N^{k-2} \otimes I)\mathrm{d}y.
\end{aligned}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Assuming now that the result holds until rank $p$ with $1 \le p \le k-2$, we prove it at rank $p+1$ thanks to analogous computations:

$$\begin{aligned} M^k &= (-1)^{p+1} \int_Y (M^{k-p} + 2\partial_j N^{k-p-1} \otimes e_j + N^{k-p-2} \otimes I) \otimes N^p \\ &\quad - N^{k-p-1} \otimes N^{p-1} \otimes I) \mathrm{d}y \\ &= (-1)^{p+1} \int_Y ((-2\partial_j N^p \otimes e_j - N^{p-1} \otimes I) \otimes N^{k-p-1} \\ &\quad + N^{k-p-2} \otimes N^p \otimes I) \mathrm{d}y \\ &= (-1)^{p+1} \int_Y ((\Delta_{yy} N^{p+1} + M^{p+1}) \otimes N^{k-p-1} + N^{k-p-2} \otimes N^p \otimes I) \mathrm{d}y. \end{aligned}$$

□

## 10.3.4 Simplifications for the tensors $\mathcal{X}^{k*}$ and $M^k$ in case of symmetries

In this last part, we analyze how the homogenized tensors $\mathcal{X}^{k*}$ and $M^k$ reduce to small number of effective coefficients when the obstacle $\eta T$ is symmetric with respect to axes of the unit cell $P$. Such results are classical in the theory of homogenization; our methodology follows, e. g., Section 6 in [13].

In all what follows, we denote by $S := (S_{ij})_{1 \le i,j \le d}$ an arbitrary orthogonal symmetry (satisfying $S = S^T$ and $SS = I$). We will specify $S$ in Corollary 10.3.3 to either of the following cell symmetries:

- for $1 \le l \le d$, $S^l$ denotes the symmetry with respect to the hyperplane orthogonal to $\boldsymbol{e}_l$:

$$S^l := I - 2\boldsymbol{e}_l \boldsymbol{e}_l^T;$$

- for $1 \le m \ne l \le d$, $S^{lm}$ denotes the symmetry with respect to the diagonal hyperplane that is orthogonal to $\boldsymbol{e}_l - \boldsymbol{e}_m$:

$$S^{lm} := I - \boldsymbol{e}_l \boldsymbol{e}_l^T - \boldsymbol{e}_m \boldsymbol{e}_m^T + \boldsymbol{e}_l \boldsymbol{e}_m^T + \boldsymbol{e}_m \boldsymbol{e}_l^T.$$

Recall that the Laplace operator is invariant under such orthogonal symmetries $S$: for any smooth scalar field $\mathcal{X}$,

$$-\Delta(\mathcal{X} \circ S) = -(\Delta \mathcal{X}) \circ S.$$

**Proposition 10.3.9.** *If the cell $Y = P \setminus \eta T$ is invariant with respect to a symmetry $S$,* i. e., *$S(Y) = Y$, then we have the following identity for the components of the solutions $\mathcal{X}^k$ to*

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*the cell problem* (10.31)*:*

$$\mathcal{X}^k_{i_1\dots i_k} \circ S = S_{i_1 j_1} \dots S_{i_k j_k} \mathcal{X}^k_{j_1\dots j_k},$$

*where we recall the implicit summation convention over the repeated indices* $j_1 \dots j_k$.

*Proof.* The result is proved by induction on $k$. For $k = 0$, we have

$$-\Delta_{yy}(\mathcal{X}^0 \circ S) = 1 \circ S = 1,$$

and the symmetry of $Y$ implies that $\mathcal{X}^0 \circ S$ also satisfies the boundary conditions (10.31). This implies $\mathcal{X}^0 \circ S = \mathcal{X}^0$. For $k = 1$, we write

$$-\Delta_{yy}(\mathcal{X}^1_{i_1} \circ S) = 2(\partial_{i_1}\mathcal{X}^0) \circ S = 2\partial_{j_1}(\mathcal{X}^0 \circ S) S_{i_1 j_1} = 2\partial_{j_1}\mathcal{X}^0 S_{i_1 j_1},$$

which similarly implies $\mathcal{X}^1_{i_1} \circ S = S_{i_1 j_1}\mathcal{X}^1_{j_1}$. Finally, if the result holds till rank $k+1$ with $k \geq 0$, then

$$\begin{aligned}
-\Delta_{yy}(\mathcal{X}^{k+2}_{i_1\dots i_{k+2}} \circ S) &= 2(\partial_{i_{k+2}}\mathcal{X}^{k+1}_{i_1\dots i_{k+1}}) \circ S + \delta_{i_{k+1}i_{k+2}}\mathcal{X}^k_{i_1\dots i_k} \circ S \\
&= 2S_{i_{k+2}j_{k+2}}\partial_{j_{k+2}}(\mathcal{X}^{k+1}_{i_1\dots i_{k+1}} \circ S) + S_{i_{k+1}j_{k+1}}S_{i_{k+2}j_{k+2}}\delta_{j_{k+1}j_{k+2}}\mathcal{X}^k_{i_1\dots i_k} \circ S \\
&= -S_{i_1 j_1} \dots S_{i_k j_{k+2}}\Delta_{yy}\mathcal{X}^{k+2}_{j_1\dots j_{k+2}},
\end{aligned}$$

whence the result at rank $k+2$. □

**Corollary 10.3.2.** *If the cell* $Y = P \setminus \eta T$ *is invariant with respect to a symmetry* $S$*, then the components of the tensors* $\mathcal{X}^{k*}$ *and* $M^k$ *of, respectively,* (10.35) *and* (10.38) *satisfy*

$$\mathcal{X}^{k*}_{i_1\dots i_k} = S_{i_1 j_1} \dots S_{i_k j_k}\mathcal{X}^{k*}_{j_1\dots j_k}, \tag{10.48}$$

$$M^k_{i_1\dots i_k} = S_{i_1 j_1} \dots S_{i_k j_k} M^k_{j_1\dots j_k}, \tag{10.49}$$

*where we recall the implicit summation over the repeated indices* $j_1 \dots j_k$.

*Proof.* Equality (10.48) results from the previous proposition and from the following change of variables:

$$\mathcal{X}^{k*}_{i_1\dots i_k} = \int_Y \mathcal{X}^k_{i_1\dots i_k}\,\mathrm{d}y = \int_Y \mathcal{X}^k_{i_1\dots i_k} \circ S\mathrm{d}y.$$

Equality (10.49) can be obtained by using equality (10.48) in formula (10.41). □

**Corollary 10.3.3.**

1. *If the cell* $Y$ *is symmetric with respect to all cell axes* $\boldsymbol{e}_l$*, i.e.,* $S^l(Y) = Y$ *for any* $1 \leq l \leq d$*, then*

$$\mathcal{X}^{k*}_{i_1\dots i_k} = 0 \quad \textit{and} \quad M^k_{i_1\dots i_k} = 0$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*whenever there exists a number $r$ occurring with odd multiplicity in the indices $i_1 \dots i_k$, i.e., whenever*

$$\exists r \in \{1, \dots, d\}, \quad \mathrm{Card}\{j \in \{1, \dots, k\} \mid i_j = r\} \text{ is odd.}$$

2. *If the cell $Y$ is symmetric with respect to all diagonal axes $(\boldsymbol{e}_l - \boldsymbol{e}_m)$, i.e., $S^{l,m}(Y) = Y$ for any $1 \le l < m \le d$, then for any permutation $\sigma \in \mathfrak{S}_d$,*

$$\mathcal{X}^{k*}_{\sigma(i_1)\dots\sigma(i_k)} = \mathcal{X}^{k*}_{i_1\dots i_k},$$
$$M^k_{\sigma(i_1)\dots\sigma(i_k)} = M^k_{i_1\dots i_k}.$$

*Proof.*

1. The symmetry $S^l$ is a diagonal matrix satisfying $S^l \boldsymbol{e}_l = -\boldsymbol{e}_l$ and $S^l \boldsymbol{e}_q = \boldsymbol{e}_q$ for $q \neq l$. Hence, replacing $S$ by $S^l$ in (10.48), we have

$$\mathcal{X}^{k*}_{i_1\dots i_k} = (-1)^{\delta_{i_1 l} + \dots + \delta_{i_k l}} \mathcal{X}^{k*}_{i_1\dots i_k},$$

   which implies the result.
2. Applying (10.48) to the symmetry $S^{l,m}$ yields the result for $\sigma = \tau$, where $\tau$ is the transposition exchanging $l$ and $m$. Since this holds for any transposition $\tau \in \mathfrak{S}_d$, this implies the statement for any permutation $\sigma \in \mathfrak{S}_d$. □

Let us illustrate how the previous corollary reads for the tensors $M^2$ and $M^4$:

- if $Y$ is symmetric with respect to the cell axes $(\boldsymbol{e}_l)_{1 \le l \le d}$, then only the coefficients $M^2_{ii}, M^4_{iijj}, M^4_{iiii}$ with $1 \le i, j \le d$ and $i \neq j$ are nonzeros (in particular, $M^2$ is diagonal).
- if, in addition, $Y$ is symmetric with respect to the hyperplane orthogonal to $\boldsymbol{e}_l - \boldsymbol{e}_m$, then these coefficients do not depend on the values of the distinct indices $i$ and $j$. As a result, $M^2$ is a multiple of the identity, and $M^4$ reduces to two effective coefficients: there exist constants $\alpha, \beta, \nu \in \mathbb{R}$ such that

$$M^2 \cdot \nabla^2 = \nu \Delta \quad \text{and} \quad M^4 \cdot \nabla^4 = \alpha \sum_{i=1}^{d} \partial^4_{iiii} + \beta \sum_{1 \le i \neq j \le d} \partial^4_{iijj}.$$

## 10.4 Homogenized equations of order $2K + 2$: tensors $\mathbb{D}^{2k}_K$

This section details steps (4) and (5) of Section 10.2.2 concerned with the process of truncating the infinite-order homogenized equation (10.23) so as to obtain well-posed effective models of finite order. Recall that this process is needed because equation (10.10) is in general ill-posed, since the tensors $M^k$ have no any particular sign (in view of formula (10.41)).

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

In Section 10.4.1, we introduce the main technical results, which allow us to derive error estimates. More particularly, we show in Section 10.4.1 that for any integer $K' \in \mathbb{N}$, any family of nonoscillating functions $(v_\epsilon^*)_{\epsilon>0}$ yields an error estimate of order $O(\epsilon^{K'+3})$, provided that (i) $v_\epsilon^*$ is of order $O(\epsilon^2)$ and (ii) $v_\epsilon^*$ solves the infinite-order homogenized equation (10.9) up to a remainder of order $O(\epsilon^{K'+1})$:

$$\sum_{k=0}^{K'} \epsilon^{k-2} M^k \cdot \nabla^k v_\epsilon^* = f + O(\epsilon^{K'+1}). \tag{10.50}$$

In particular, this result reminds us that higher-order models are generally not unique; they differ by the choice of extra differential operators of order greater than $K'$, which turn equation (10.50) into a well-posed model.

Leaving momentarily these considerations aside, we propose in Section 10.4.2 a "variational" method inspired by [23, 52], which allows us to construct a well-posed effective model (10.6) of order $2K+2$ for any $K \in \mathbb{N}$. The procedure relies on an energy minimization principle based on the criminal ansatz (10.11); the coefficients $\mathbb{D}_K^{2k}$ are inferred from an effective energy $J_K^*$ (Definition 10.4.2) and are a priori distinct from the tensors $M^{2k}$. These properties enable us to establish that the obtained model is *elliptic* (in particular, well-posed) and hence amenable to numerical computations.

Finally, we obtain in Section 10.4.3 that, surprisingly, $\mathbb{D}_K^{2k} = M^{2k}$ for any $0 \leq 2k \leq K$, which implies that equation (10.50) is satisfied by the solution $v_\epsilon^* \equiv v_{\epsilon,K}^*$ with $K' = 2K+1$ (recall that $M^{2k+1} = 0$ for any $k \in \mathbb{N}$ from Corollary 10.3.1). The error estimate (10.14) of order $O(\epsilon^{2K+4})$ follows in Corollary 10.4.2.

## 10.4.1 Sufficient conditions under which an effective solution yields higher-order approximations

The main result of this part is Proposition 10.4.1, where we provide sufficient conditions under which a sequence of macroscopic functions $v_\epsilon^* \in \mathcal{C}^\infty(D)$ (depending on $\epsilon$, $K$, and $f$) yields a high-order approximation of $u_\epsilon$. The proof is based on the properties of the tensors $N^k$ stated in Proposition 10.3.8 and the next three results.

**Lemma 10.4.1** (see, e. g., Lions (1981) [41]). *There exists a constant $C$ independent of $\epsilon$ such that for any $\phi \in H^1(D_\epsilon)$ satisfying $\phi = 0$ on the boundary $\partial\omega_\epsilon$ of the holes, we have the following Poincaré inequality:*

$$\|\phi\|_{L^2(D_\epsilon)} \leq C\epsilon \|\nabla\phi\|_{L^2(D_\epsilon,\mathbb{R}^d)}.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Corollary 10.4.1.** *For any $h \in L^2(D_\epsilon)$, let $r_\epsilon \in H^1(D_\epsilon)$ be the unique solution to the Poisson problem*

$$\begin{cases} -\Delta r_\epsilon = h \quad \text{in } D_\epsilon, \\ r_\epsilon = 0 \quad \text{on } \partial\omega_\epsilon, \\ r_\epsilon \text{ is } D\text{-periodic.} \end{cases}$$

*There exists a constant $C$ independent of $\epsilon$ and $h$ such that*

$$\|r_\epsilon\|_{L^2(D_\epsilon)} + \epsilon\|\nabla r_\epsilon\|_{L^2(D_\epsilon,\mathbb{R}^d)} \le C\|h\|_{L^2(D_\epsilon)}\epsilon^2. \tag{10.51}$$

*Proof.* This result is classical; it is a consequence of Lemma 10.4.1 and of the energy estimate

$$\int_{D_\epsilon} |\nabla r_\epsilon|^2 \mathrm{d}x = \int_{D_\epsilon} h r_\epsilon \mathrm{d}x \le \|h\|_{L^2(D_\epsilon)}\|r_\epsilon\|_{L^2(D_\epsilon)} \le C\epsilon\|h\|_{L^2(D_\epsilon)}\|\nabla r_\epsilon\|_{L^2(D_\epsilon,\mathbb{R}^d)},$$

which implies $\|\nabla r_\epsilon\|_{L^2(D_\epsilon,\mathbb{R}^d)} \le C\|h\|_{L^2(D_\epsilon)}\epsilon$ and then estimate (10.51). □

**Lemma 10.4.2.** *Assume that the boundary of obstacle $\eta T$ is smooth. Then, for any $k \in \mathbb{N}$, the tensor $\mathcal{X}^k$ is well-defined and is smooth, namely, $\mathcal{X}^k \in \mathcal{C}^\infty(\overline{Y})$. In particular, $\mathcal{X}^k \in L^\infty(Y) \cap H^1(Y)$.*

*Proof.* Since the constant function 1 is smooth, standard regularity theory for the Laplace operator $-\Delta_{yy}$ (see [22, 32, 35]) implies $\mathcal{X}^0 \in \mathcal{C}^\infty(\overline{Y})$. The result follows by induction by repeating this argument to $\mathcal{X}^1$ and $\mathcal{X}^{k+2}$ for any $k \ge 0$. □

**Proposition 10.4.1.** *Let $v_\epsilon^* \in \mathcal{C}^\infty(D)$ be a $D$-periodic function depending on $\epsilon$ (and, possibly, on $K'$ and $f$) satisfying the following two hypotheses:*

1. *for any $m \in \mathbb{N}$, there exists a constant $C_{K',m}(f)$ depending only on $m$, $K'$, and $f \in \mathcal{C}^\infty(D)$ such that*

$$\|v_\epsilon^*\|_{H^m(D)} \le C_{K',m}(f)\epsilon^2. \tag{10.52}$$

2. *$v_\epsilon^*$ solves the infinite-order homogenized equation* (10.9) *up to a remainder of order $O(\epsilon^{K'+1})$:*

$$\left\| \sum_{k=0}^{K'} \epsilon^{k-2} M^k \cdot \nabla^k v_\epsilon^* - f \right\|_{L^2(D)} \le C_{K'}(f)\epsilon^{K'+1}. \tag{10.53}$$

*Then the reconstructed function $W_{\epsilon,K'}(v_\epsilon^*)$ of* (10.12) *approximates the solution $u_\epsilon$ of the perforated Poisson problem* (10.1) *at order $O(\epsilon^{K'+3})$, viz. there exists a constant $C_{K'}(f)$ independent of $\epsilon$ such that*

$$\|u_\epsilon - W_{\epsilon,K'}(v_\epsilon^*)\|_{L^2(D_\epsilon)} + \epsilon\|\nabla(u_\epsilon - W_{\epsilon,K'}(v_\epsilon^*))\|_{L^2(D_\epsilon,\mathbb{R}^d)} \le C_{K'}(f)\epsilon^{K'+3}.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*Proof.* Let us compute

$$\begin{aligned}
-\Delta W_{\epsilon,K'}(v_\epsilon^*) &= \sum_{k=0}^{K'} \epsilon^{k-2}(-\Delta N^k - 2\partial_l N^{k-1}\otimes e_l - N^{k-2}\otimes I)(\cdot/\epsilon)\cdot\nabla^k v_\epsilon^* \\
&\quad - \epsilon^{K'-1}(2\partial_l N^{K'}\otimes e_l + N^{K'-1}\otimes I)(\cdot/\epsilon)\cdot\nabla^{K'+1}v_\epsilon^* \\
&\quad - \epsilon^{K'}N^{K'}(\cdot/\epsilon)\otimes I\cdot\nabla^{K'+2}v_\epsilon^* \\
&= \sum_{k=0}^{K'} \epsilon^{k-2}M^k\cdot\nabla^k v_\epsilon^* - \epsilon^{K'-1}(2\partial_l N^{K'}\otimes e_l + N^{K'-1}\otimes I)(\cdot/\epsilon)\cdot\nabla^{K'+1}v_\epsilon^* \\
&\quad - \epsilon^{K'}N^{K'}(\cdot/\epsilon)\otimes I\cdot\nabla^{K'+2}v_\epsilon^*,
\end{aligned}$$

where we have used (10.45) to obtain the second equality. Since the functions $(N^k)_{k\in\mathbb{N}}$ are smooth (Lemma 10.4.2 and (10.44)), assumption (10.52) implies that the last two terms are lower than $\epsilon^{K'+1}$:

$$\begin{aligned}
\left\|\epsilon^{K'-1}(2\partial_l N^{K'}\otimes e_l + N^{K'-1}\otimes I)(\cdot/\epsilon)\cdot\nabla^{K'+1}v_\epsilon^*\right\| &\le C_{K'}(f)\epsilon^{K'+1}, \\
\left\|\epsilon^{K'}N^{K'}(\cdot/\epsilon)\otimes I\cdot\nabla^{K'+2}v_\epsilon^*\right\|_{L^2(D_\epsilon)} &\le C_{K'}(f)\epsilon^{K'+1}.
\end{aligned}$$

Using now assumption (10.53) and applying Corollary 10.4.1 to $r_\epsilon := u_\epsilon - W_{\epsilon,K'}(v_\epsilon^*)$ yields the result. □

## 10.4.2 Construction of a well-posed higher-order effective model by means of a variational principle

Leaving momentarily aside the result of Proposition 10.4.1, we now detail the construction of our effective model (10.6) of finite order inspired by the works [15, 23, 52]. The construction of the coefficients $\mathbb{D}_K^{2k}$ from an effective energy is exposed in Section 10.4.2.1, and the well-posedness of the effective model is established in Section 10.4.2.2.

### 10.4.2.1 The method of Smyshlyaev and Cherednichenko

According to the ideas outlined in step (4) of Section 10.2.2, we consider truncations $W_{\epsilon,K}(v)$ of the "criminal" ansatz (10.25) of the form

$$W_{\epsilon,K}(v)(x) := \sum_{k=0}^{K} \epsilon^k N^k(x/\epsilon)\cdot\nabla^k v(x), \quad x\in D, \tag{10.54}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

where we seek a function $v \in H^{K+1}(D)$ that does not depend on the fast variable $x/\epsilon$ and that approximates the formal homogenized average $u_\epsilon^*$ of (10.22). The tensors $N^k(y)$ are extended by 0 in $P \setminus Y$ for (10.54) to make sense in $D \setminus D_\epsilon$.

For any $u \in H^1(D)$ and $f \in L^2(D)$, we denote by $J(u,f)$ the energy

$$J(u,f) := \int_D \left( \frac{1}{2}|\nabla u|^2 - fu \right) \mathrm{d}x.$$

Following the lines of point (4) of Section 10.2.2, we consider problem (10.28) of finding a minimizer of $v \mapsto J(W_{\epsilon,K}(v), f)$. The key step of the strategy is to eliminate the fast variable $x/\epsilon$ in $J(W_{\epsilon,K}(v), f)$ so as to obtain an effective energy $J_K^*(v, f, \epsilon) \simeq J(W_{\epsilon,K}(v), f)$, which does not involve oscillating functions. The main technical tool that allows us to perform this operation is the following classical lemma of two-scale convergence (see, e. g., Appendix C of [52] or [6]).

**Lemma 10.4.3.** *Let $\phi$ be a $P = [0,1]^d$-periodic function, and let $f \in \mathcal{C}^\infty(D)$ be a smooth $D$-periodic function. Then for any $k \in \mathbb{N}$ arbitrarily large, we have the following inequality:*

$$\left| \int_D f(x)\phi(x/\epsilon)\mathrm{d}x - \int_D \int_P f(x)\phi(y)\mathrm{d}y\mathrm{d}x \right| \leq \frac{L^{d/2}}{|2\pi|^p} \left\| \phi - \int_P \phi \mathrm{d}y \right\|_{L^2(P)} \|f\|_{H^p(D)} \epsilon^p.$$

*Proof.* See Lemma 7.3 in [33] for a proof of this statement. □

Before providing the definition of $J_K^*$ based on the application of Lemma 10.4.3, we introduce several additional tensors that arise in the averaging process.

**Definition 10.4.1** (Tensors $\mathbb{B}_K^{l,m}$). For any $K \in \mathbb{N}$, $1 \leq j \leq d$, and $0 \leq k \leq K+1$, let $\widetilde{N}_j^k(y)$ (with implicit dependence with respect to $K$) be the $k$th-order tensor defined by

$$\widetilde{N}_j^k(y) = \begin{cases} \partial_j N^0(y) & \text{if } k = 0, \\ \partial_j N^k(y) + N^{k-1}(y) \otimes e_j & \text{if } 1 \leq k \leq K, \\ N^K(y) \otimes e_j & \text{if } k = K+1. \end{cases} \tag{10.55}$$

We define a family of constant bilinear tensors $\mathbb{B}_K^{l,m}$ of order $l+m$ by the formula

$$\mathbb{B}_K^{l,m} := \int_Y \widetilde{N}_j^l(y) \otimes \widetilde{N}_j^m(y)\mathrm{d}y \quad \text{for } 0 \leq l, m \leq K+1, \tag{10.56}$$

where the Einstein summation convention is still assumed over the repeated subscript indices $1 \leq j \leq d$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Definition 10.4.2** (Approximate energy $J_K^*$). For any $f \in L^2(D)$ and periodic function $v \in H^{K+1}(D)$, we define

$$J_K^*(v,f,\epsilon) := \int_D \left( \frac{1}{2} \sum_{l,m=0}^{K+1} \epsilon^{l+m-2} \mathbb{B}_K^{l,m} \nabla^l v \nabla^m v - fv \right) \mathrm{d}x, \tag{10.57}$$

where we recall (10.19) for the definition of $\mathbb{B}_K^{l,m} \nabla^l v \nabla^m v$.

The definition of the energy $J_K^*(v,f,\epsilon)$ is motivated by the following asymptotics, provided by Lemma 10.4.3, which holds for any $p \geq 0$ arbitrarily large:

$$J(W_{\epsilon,K}(v),f) = J_K^*(v,f,\epsilon) + o(\epsilon^p).$$

More precisely, we have the following result.

**Proposition 10.4.2.** *Let $f \in \mathcal{C}^\infty(D)$ be $D$-periodic. Let $v \in \mathcal{C}^\infty(D)$ be a smooth $D$-periodic function, and let $W_{\epsilon,K}(v) \in \mathcal{C}^\infty(D_\epsilon)$ be the truncated ansatz of the form* (10.54)*. We have the following estimate for $p \geq 0$ arbitrarily large:*

$$\left| J(W_{\epsilon,K}(v),f) - J_K^*(v,f,\epsilon) \right| \leq C_{K,p} \left( \|v\|^2_{H^{p+2}(D)} + \|f\|^2_{H^p(D)} \right) \epsilon^p$$

*for a constant $C_{K,p}$ depending only on $p$ and $K$ (and $\eta$).*

*Proof.* For any $1 \leq j \leq d$, the partial derivative $\partial_{x_j} W_{\epsilon,K}(v)$ reads

$$\begin{aligned} \partial_{x_j} W_{\epsilon,K}(v) &= \sum_{k=0}^{K} \left( \epsilon^{k-1} \partial_{y_j} N^k(\cdot/\epsilon) \cdot \nabla^k v + \epsilon^k N^k(\cdot/\epsilon) \otimes e_j \cdot \nabla^{k+1} v \right) \\ &= \sum_{i=0}^{K+1} \epsilon^{k-1} \widetilde{N}_j^k(\cdot/\epsilon) \cdot \nabla^k v \end{aligned}$$

by definition (10.55) of the tensors $\widetilde{N}_j^k$. Then the computation of the energy $J(W_{\epsilon,K}(v),f)$ yields

$$\begin{aligned} J(W_{\epsilon,K}(v),f) = &\int_D \left( \frac{1}{2} \sum_{l,m=0}^{K+1} \epsilon^{l+m-2} (\widetilde{N}_j^l(\cdot/\epsilon) \cdot \nabla^l v)(\widetilde{N}_j^m(\cdot/\epsilon) \cdot \nabla^m v) \right) \mathrm{d}x \\ &- \int_D \sum_{l=0}^{K} \epsilon^l (N^l(\cdot/\epsilon) \cdot \nabla^l v) f) \mathrm{d}x. \end{aligned}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

The result follows by estimating both terms after applying Lemma 10.4.3: for any $0 \le l$, $m \le K+1$,

$$\begin{aligned}
&\left|\int_D \epsilon^{l+m-2}((\widetilde{N}_j^l(\cdot/\epsilon)\cdot\nabla^l v)(\widetilde{N}_j^m(\cdot/\epsilon)\cdot\nabla^m v) - \mathbb{B}_K^{l,m}\nabla^l v\nabla^m v)\mathrm{d}x\right| \\
&\quad \le C_p\epsilon^p \left\|\nabla^l v \otimes \nabla^m v\right\|_{H^{p-(l+m-2)}(D,\mathbb{R}^{d^{l+m}})} \\
&\quad \le C_p'\epsilon^p\left(\left\|\nabla^l v\right\|^2_{H^{p-(l+m-2)}(D,\mathbb{R}^{d^l})} + \left\|\nabla^m v\right\|^2_{H^{p-(l+m-2)}(D,\mathbb{R}^{d^m})}\right) \\
&\quad \le C_p''\epsilon^p \|v\|^2_{H^{p+2}(D)}, \\
\forall 0 \le l \le K, \quad &\left|\int_D \epsilon^l(N^l(\cdot/\epsilon)\cdot\nabla^l v f - fv)\mathrm{d}x\right| \le C_p\epsilon^p\left\|f\nabla^l v\right\|_{H^{p-l}(D,\mathbb{R}^{d^l})} \\
&\qquad\qquad \le C_p'\epsilon^p\left(\|f\|^2_{H^{p-l}(D)} + \|v\|^2_{H^p(D)}\right) \\
&\qquad\qquad \le C_p'\epsilon^p\left(\|f\|^2_{H^p(D)} + \|v\|^2_{H^p(D)}\right),
\end{aligned}$$

where we used that $N^l$ is of average 1 if $l = 0$ and 0 otherwise (item 1 of Proposition 10.3.8). □

The approximate energy (10.57) is used (instead of $J(W_{\epsilon,K}(v), f)$ in (10.28)) to construct our higher-order homogenized model.

**Definition 10.4.3.** For any $K \in \mathbb{N}$, we call the homogenized equation of order $2K+2$ the Euler–Lagrange equation associated with the minimization problem

$$\begin{aligned}
\min_{v\in H^{K+1}(D)} \quad & J_K^*(v, f, \epsilon) \\
\text{s. t.} \quad & v \text{ is } D\text{-periodic.}
\end{aligned} \tag{10.58}$$

This equation reads explicitly in terms of a higher-order homogenized solution $v_K^* \in H^{K+1}(D)$ as

$$\begin{cases} \sum_{k=0}^{K+1} \epsilon^{2k-2}\mathbb{D}_K^{2k}\cdot\nabla^{2k} v_{\epsilon,K}^* = f, \\ v_{\epsilon,K}^* \text{ is } D\text{-periodic,} \end{cases} \tag{10.59}$$

where the constant tensors $\mathbb{D}_K^{2k}$ are defined for any $0 \le 2k \le 2K+2$ as

$$\mathbb{D}_K^{2k} := \sum_{l=0}^{2k}(-1)^l \mathbb{B}_K^{l,2k-l}, \tag{10.60}$$

assuming the convention $\mathbb{B}_K^{l,m} = 0$ for any $l > K+1$ or $m > K+1$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*Proof.* Let us detail slightly the derivation of (10.59). The Euler–Lagrange equation of (10.58) reads, after an integration by parts,

$$\sum_{l,m=0}^{K+1} \epsilon^{l+m-2} \frac{1}{2}\left((-1)^m \mathbb{B}_K^{l,m} + (-1)^l \mathbb{B}_K^{m,l}\right) \nabla^{l+m} v_{\epsilon,K}^* = f. \tag{10.61}$$

Since $\mathbb{B}_K^{l,m} = \mathbb{B}_K^{m,l}$ and $(-1)^l + (-1)^m$ vanishes when $l$ and $m$ are not of the same parity, only terms such that $l + m$ is even are possibly not zero in the above equation. Hence, equation (10.61) rewrites as equation (10.59) with

$$\mathbb{D}_K^{2k} = \sum_{l+m=2k} \frac{1}{2}\left((-1)^l + (-1)^m\right) \mathbb{B}_K^{l,m} = \sum_{l=0}^{2k} \frac{1}{2}\left((-1)^l + (-1)^{2k-l}\right) \mathbb{B}_K^{l,2k-l}$$

for any $0 \leq 2k \leq 2K + 2$, which eventually yields the desired expression (10.60). □

**Remark 10.4.1.** As announced in the introduction, equation (10.59) turns out to be a simple correction of equation (10.10); see Proposition 10.4.4.

#### 10.4.2.2 Well-posedness of the homogenized model of order $2K + 2$

We now establish the well-posedness of the high-order homogenized model (10.59). More precisely, we prove its ellipticity, which implies the existence and uniqueness of the effective solution $v_{\epsilon,K}^*$. Before stating the result, let us stress the following observation, which is obvious but a somewhat important consequence of definition (10.56).

**Lemma 10.4.4.** *The dominant tensor $\mathbb{B}_K^{K+1,K+1}$ is symmetric and nonnegative.*

**Proposition 10.4.3.** *Assume further that the dominant tensor $\mathbb{B}_K^{K+1,K+1}$ is nondegenerate, viz. there exists a constant $\nu > 0$ such that*

$$\forall \boldsymbol{\xi} = \xi_{i_1 \dots i_{K+1}} \in \mathbb{R}^{d^{K+1}}, \quad \mathbb{B}_K^{K+1,K+1} \cdot \boldsymbol{\xi}\boldsymbol{\xi} \geq \nu |\boldsymbol{\xi}|^2. \tag{10.62}$$

*Then equation* (10.59) *is elliptic, and there exists a unique solution $v_K^* \in H^{K+1}(D)$ to the homogenized equation* (10.59) *of order $2K + 2$.*

*Proof.* Let us consider the space $V_K := \{v \in H^{K+1}(D) \mid v \text{ is } D\text{-periodic}\}$ and introduce the respective bilinear and linear forms $a : V_K \times V_K \to \mathbb{R}$ and $b : V_K \to \mathbb{R}$ defined for any $v \in V_K$ by

$$a(v, v) = \int_D \sum_{l,m=0}^{K+1} \epsilon^{l+m-2} \mathbb{B}_K^{l,m} \nabla^l v \nabla^m v \mathrm{d}x,$$

$$b(v) = \int_D f v \mathrm{d}x.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

The homogenized equation (10.59) reduces to the following variational problem:

$$\text{find } v^*_{\epsilon,K} \in V_K \text{ such that } \forall v \in V_K,\ a(v^*_{\epsilon,K}, v) = b(v). \tag{10.63}$$

From there we could directly rely on the theory of Fredholm operators [42] to conclude the existence of a solution $v^*_{\epsilon,K}$. However, we are going to show that $a$ is coercive (meaning that equation (10.59) is elliptic), which will allow us to apply the Lax–Milgram theorem [31].

Under the nondegeneracy assumption (10.62), it is readily obtained that there exists a constant $C_\epsilon$ (depending on $\epsilon$) such that for any $v \in V_K(D)$,

$$a(v,v) \geq (\epsilon^{2K} v)\|\nabla^{K+1} v\|^2_{L^2(D,\mathbb{R}^d)} + \epsilon^{-2} M^0 \|v\|^2_{L^2(D)} - C_\epsilon \|v\|_{H^{K+1}(D)} \|v\|_{H^K(D)}.$$

Recalling that $M^0 > 0$ and applying the Young inequality

$$\forall x, y \in \mathbb{R}, \quad |xy| \leq \frac{x^2}{2\gamma} + \frac{\gamma y^2}{2},$$

for a sufficiently small $\gamma > 0$, we obtain the existence of two constants $\alpha_{\epsilon,K} > 0$ and $\beta_{\epsilon,K} > 0$ (that depend on $\epsilon$ and $K$) such that

$$\forall v \in V_K(D), \quad a(v,v) \geq \alpha_{\epsilon,K} \|v\|^2_{H^{K+1}(D)} - \beta_{\epsilon,K} \|v\|^2_{H^K(D)}. \tag{10.64}$$

Furthermore, equation (10.56), together with the proof of Proposition 10.4.2, allows us to rewrite $a(v,v)$ as

$$a(v,v) = \int_D \int_Y \left| (\epsilon^{-1}\nabla_y + \nabla_x) \left( \sum_{k=0}^{K} \epsilon^k N^k(y) \cdot \nabla^k v(x) \right) \right|^2 \mathrm{d}x. \tag{10.65}$$

Then $\int_D \int_Y u(x,y)^2 \mathrm{d}y\mathrm{d}x \geq \int_D |\int_Y u(x,y)\mathrm{d}y|^2 \mathrm{d}x$ and item 1 of Proposition 10.3.8 imply the following inequality:

$$\forall v \in V_K, \quad a(v,v) \geq \|\nabla v\|^2_{L^2(D,\mathbb{R}^d)}. \tag{10.66}$$

We now prove that equations (10.64) and (10.66) together imply the coercivity of $a$ on the space $V_K$, that is, we claim there exists a constant $c_{\epsilon,K} > 0$ such that

$$\forall v \in V_K, \quad a(v,v) \geq c_{\epsilon,K} \|v\|^2_{H^{K+1}(D)}. \tag{10.67}$$

Assume the contrary is true. Then we can find a sequence $(v_n)$ of functions satisfying $\|v_n\|_{H^{K+1}(D)} = 1$ and such that $a(v_n, v_n) \to 0$. Up to extracting a relevant subsequence, we may assume that $v_n \rightharpoonup v$ weakly in $H^{K+1}(D)$ and $v_n \to v$ strongly in $H^K(D)$. Then the

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

polarization identity, together with (10.64) and the positivity of $a$, allows us to show that $(v_n)$ is a Cauchy sequence in $V_K$:

$$\begin{aligned}\forall p,q\in\mathbb{N},\quad \alpha_{\epsilon,K}\|v_p-v_q\|^2_{H^{K+1}(D)} &\le a(v_p-v_q,v_p-v_q)+\beta_{\epsilon,K}\|v_p-v_q\|^2_{H^K(D)}\\ &= 2a(v_p,v_p)+2a(v_q,v_q)-a(v_p+v_q,v_p+v_q)\\ &\quad+\beta_{\epsilon,K}\|v_p-v_q\|^2_{H^K(D)}\\ &\le 2a(v_p,v_p)+2a(v_q,v_q)+\beta_{\epsilon,K}\|v_p-v_q\|^2_{H^K(D)}\xrightarrow{p,q\to\infty}0.\end{aligned}$$

Therefore $v_n\to v$ strongly in $V_K$. Using the continuity of $a$, we infer that $a(v,v)=\lim_{n\to+\infty}a(v_n,v_n)=0$. Then property (10.66) yields that $v$ is a constant. Therefore $0=a(v,v)=\epsilon^{-2}M^0\|v\|^2_{L^2(D)}$, which implies $v=0$. This is in contradiction with the fact that $\|v_n\|_{H^{K+1}(D)}=1$ for any $n\ge 0$ and the strong convergence of $(v_n)$ in $H^{K+1}(D)$.

Finally, coercivity (10.67) and the continuity of $a$ and $b$ over $V_K$ ensure that all the assumptions of the Lax–Milgram theorem are fulfilled, which yields the existence and uniqueness of a solution to problem (10.63). □

**Remark 10.4.2.** It is always possible to add to $\mathbb{D}_K^{2K+2}$ a small perturbation making equation (10.62) satisfied while keeping an "admissible" higher-order homogenized equation. Indeed, since the other $2K+1$ coefficients are kept unaffected, the error estimate provided by Proposition 10.4.1 and Corollary 10.4.2 below remain valid. Let us remark, however, that this nondegeneracy condition is automatically fulfilled for any shape of obstacle $\eta T$ when $K=0$ because it is easily shown that $\mathbb{D}_0^2=-(M^0)^2\int_Y|\mathcal{X}^0(y)|^2\mathrm{d}y>0$ (see (10.69)). In the general case $K\ge 1$, it could fail to be satisfied for particular obstacle shapes (e. g., in case of invariance of $\eta T$ along some of the cell axes).

### 10.4.3 Asymptotic surprise: $\mathbb{D}_K^{2k}=M^{2k}$; error estimates for the homogenized model of order $2K+2$

We terminate this section by verifying the assumptions of Proposition 10.4.1, which ensure the validity of the error estimate (10.14) claimed in the introduction. The next proposition establishes item 1.

**Lemma 10.4.5.** *Assume the nondegeneracy condition* (10.62). *The solution* $v^*_{\epsilon,K}$ *of* (10.59) *belongs to* $\mathcal{C}^\infty(D)$, *and for any* $m\in\mathbb{N}$, *there exists a constant* $C_{m,K}$ *that does not depend on* $\epsilon$ *such that*

$$\left\|v^*_{\epsilon,K}\right\|_{H^m(D)}\le C_{m,K}\|f\|_{H^m(D)}\epsilon^2. \tag{10.68}$$

*Proof.* We solve equation 10.59 explicitly with Fourier expansions in the periodic domain $D=[0,L]^d$. Let $\hat{f}(\xi)$ be the Fourier coefficients of $f$, and let $c(\xi,\epsilon)$ be the symbol

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

of the differential operator (10.59), namely:

$$\forall \xi \in \mathbb{Z}^d,\ \forall \epsilon > 0, \quad c(\xi, \epsilon) = \sum_{k=0}^{K+1} (-1)^k |2\pi/L|^{2k} \epsilon^{2k-2} \mathbb{D}_K^{2k} \cdot \xi^{2k},$$

where $\xi^0 = 1$ by convention, and where we have used the short-hand notation $\mathbb{D}_K^{2k} \cdot \xi^{2k} := \mathbb{D}^{2k}_{i_1 \dots i_{2k}} \xi_{i_1 \dots i_{2k}}$. For $\xi \in \mathbb{Z}^d$, the Fourier coefficients $\hat{v}^*_{\epsilon,K}$ of $v^*_K$ read

$$\hat{v}^*_{\epsilon,K}(\xi) = \frac{\hat{f}(\xi)}{c(\xi, \epsilon)}.$$

Applying Parseval's identity to the bilinear form (10.65), we obtain

$$\begin{aligned} c(\xi, \epsilon) &= \int_Y \left| (\epsilon^{-1} \nabla_y + (2i\pi/L)\xi) \left( \sum_{k=0}^{K} (2i\pi/L)^k \epsilon^k N^k(y) \cdot \xi^k \right) \right|^2 \mathrm{d}y \\ &= \epsilon^{-2} \Phi(\epsilon \xi), \end{aligned}$$

where $\Phi$ is the function defined by

$$\begin{aligned} \Phi : \mathbb{R}^d &\longrightarrow \mathbb{R} \\ \lambda &\longmapsto \int_Y \left| (\nabla_y + (2i\pi/L)\lambda) \left( \sum_{k=0}^{K} (2i\pi/L)^k N^k(y) \cdot \lambda^k \right) \right|^2 \mathrm{d}y. \end{aligned}$$

To obtain (10.68), it suffices to prove the existence of some positive constant $C_K > 0$ such that $\Phi(\lambda) \geq C_K$ for any $\lambda \in \mathbb{R}^d$. The function $\Phi$ is clearly continuous nonnegative and satisfies $\Phi(\lambda) \to +\infty$ as $|\lambda| \to +\infty$ because of (10.62). Therefore it admits a minimum point $\lambda_0 \in (\mathbb{R}_+)^d$. Let us assume that $\Phi(\lambda_0) = 0$. Using the Cauchy–Schwarz inequality, we obtain

$$0 = \Phi(\lambda_0) \geq \left| \int_Y (\nabla_y + (2i\pi/L)\lambda_0) \left( \sum_{k=0}^{K} (2i\pi/L)^k N^k(y) \cdot \lambda_0^k \right) \mathrm{d}y \right|^2 = \frac{4\pi^2}{L^2} |\lambda_0|^2.$$

Therefore it must hold $\lambda_0 = 0$; however, this is not possible because

$$\Phi(0) = \int_Y \left| \nabla_y N^0(y) \right|^2 \mathrm{d}y = M^0 > 0.$$

Consequently, $\Phi(\lambda) \geq \Phi(\lambda_0) > 0$ for all $\lambda \in \mathbb{R}^d$, which concludes the proof. □

Item 2 of Proposition 10.4.1 is the object of the next result. More precisely, we prove that, surprisingly, the coefficients $\mathbb{D}_K^{2k}$ and $M^{2k}$ coincide for $0 \leq 2k \leq 2K$. A similar fact

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

was also observed for the (scalar) antiplane elasticity model considered in [52], but only for the first half $0 \le 2k \le K$ of the coefficients (and without providing a proof).

**Proposition 10.4.4.** *All the coefficients of the homogenized equation* (10.6) *of order* $2K+2$ *(defined in* (10.59)*) coincide with those of the formal infinite-order homogenized equation* (10.9)*, except the leading order one:*

$$\begin{aligned} \mathbb{D}_K^{2k} &= M^{2k} \quad \text{for any } 0 \le 2k \le 2K, \\ \mathbb{D}_K^{2K+2} &= (-1)^{K+1} \int_Y N^K \otimes N^K \otimes I \mathrm{d}y. \end{aligned} \tag{10.69}$$

*Proof.* First of all, formula (10.69) only rewrites formula (10.60). We show the following, slightly more general, result:

$$\forall 0 \le k \le 2K, \quad M^k = \sum_{l=0}^{k} (-1)^l \mathbb{B}_K^{l,k-l}, \tag{10.70}$$

which is sufficient for our purpose because of (10.60). We distinguish two cases.

1. *Case* $0 \le k \le K$. For $0 \le k, l \le K$, the coefficient $\mathbb{B}_K^{l,k-l}$ is given by (from (10.55))

$$\mathbb{B}_K^{l,k-l} = \int_Y (\partial_j N^l + N^{l-1} \otimes e_j) \otimes (\partial_j N^{k-l} + N^{k-l-1} \otimes e_j) \mathrm{d}y, \tag{10.71}$$

where we use the convention $N^{-1} = N^{-2} = 0$. After an integration by parts, we rewrite $\mathbb{B}_K^{l,k-l}$ as follows:

$$\begin{aligned} \mathbb{B}_K^{l,k-l} &= \int_Y (-\Delta N^l - 2\partial_j N^{l-1} \otimes e_j - N^{l-2} \otimes I) \otimes N^{k-l} \mathrm{d}y \\ &\quad + \int_Y (\partial_j N^l \otimes N^{k-l-1} \otimes e_j + N^{l-1} \otimes N^{k-l-1} \otimes I) \mathrm{d}y \\ &\quad + \int_Y (\partial_j N^{l-1} \otimes N^{k-l} \otimes e_j + N^{l-2} \otimes N^{k-l} \otimes I) \mathrm{d}y \\ &= \int_Y (M^l \otimes N^{k-l}) \mathrm{d}y + B^{k,l} + B^{k,l-1}, \end{aligned} \tag{10.72}$$

where $B^{k,l}$ is the $k$th-order tensor defined by

$$B^{k,l} := \int_Y (\partial_j N^l \otimes N^{k-l-1} \otimes e_j + N^{l-1} \otimes N^{k-l-1} \otimes I) \mathrm{d}y.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Using now item 1 of Proposition 10.3.8 and recognizing a telescopic series, we obtain

$$\begin{aligned}\sum_{l=0}^{k}(-1)^l\mathbb{B}_K^{l,k-l} &= (-1)^k M^k + \sum_{l=0}^{k}((-1)^l B^{k,l} - (-1)^{l-1}B^{k,l-1})\\ &= (-1)^k M^k + (-1)^k B^{k,k} - (-1)^{-1}B^{k,-1}.\end{aligned}$$

This implies (10.70) by using the facts that $M^k = 0$ when $k$ is odd and $B^{k,k} = B^{k,-1} = 0$ owing to our convention $N^{-1} = 0$.

2. *Case* $K+1 \le k \le 2K$. Equality (10.71) is valid for any $K+1 \le k \le 2K$ and $0 \le l \le k$ if we assume by convention (*in this part only*) that $N^m = 0$ whenever $m > K$, because of definition (10.56). Then equality (10.72) remains true if $0 \le l \le K$. Therefore, recognizing the same telescopic series, we are able to compute

$$\begin{aligned}\sum_{l=0}^{K}(-1)^l\mathbb{B}_K^{l,k-l} &= (-1)^K B^{k,K}\\ &= (-1)^K\int_Y(\partial_j N^K\otimes N^{k-K-1}\otimes e_j + N^{K-1}\otimes N^{k-K-1}\otimes I)\mathrm{d}y.\end{aligned}$$

Recalling that $\mathbb{B}_K^{l,m} = 0$ whenever $l > K+1$ or $m > K+1$, we eventually obtain

$$\begin{aligned}\sum_{l=0}^{k}(-1)^l\mathbb{B}_K^{l,k-l} &= \sum_{l=0}^{K}(-1)^l\mathbb{B}_K^{l,k-l} + (-1)^{K+1}\mathbb{B}_K^{K+1,k-K-1}\\ &= (-1)^K\int_Y(\partial_j N^K\otimes N^{k-K-1}\otimes e_j + N^{K-1}\otimes N^{k-K-1}\otimes I)\mathrm{d}y\\ &\quad + (-1)^{K+1}\int_Y(N^K\otimes e_j)\otimes(\partial_j N^{k-K-1} + N^{k-K-2}\otimes e_j)\mathrm{d}y\\ &= (-1)^K\int_Y((2\partial_j N^K\otimes e_j + N^{K-1}\otimes I)\otimes N^{k-K-1} - N^K\otimes N^{k-K-2}\otimes I)\mathrm{d}y\\ &= (-1)^K\int_Y((-\Delta_{yy}N^{K+1} - M^{K+1})\otimes N^{k-K-1} - N^K\otimes N^{k-K-2}\otimes I)\mathrm{d}y,\end{aligned}\tag{10.73}$$

where we have used (10.45) in the last equality. We now consider two cases:

– if $k = K+1$, then the above expression reads

$$\begin{aligned}\sum_{l=0}^{K+1}(-1)^l\mathbb{B}_K^{l,K+1-l} &= (-1)^K\int_Y((-\Delta_{yy}N^{K+1} - M^{K+1})\otimes N^0)\mathrm{d}y\\ &= (-1)^{K+1}M^{K+1} + (-1)^K\int_Y N^{K+1}\otimes M^0\mathrm{d}y\\ &= (-1)^{K+1}M^{K+1} = M^{K+1},\end{aligned}$$

where the last equality is a consequence of Corollary 10.3.1.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

– if $K+2 \leq k \leq 2K$, then (10.73) coincides with $M^k$ by using (10.47) (with $p = K+1$). □

**Remark 10.4.3.** As is highlighted by the proof Proposition 10.4.4, the "surprising" fact that $\mathbb{D}_K^{2k} = M^{2k}$ even for $K+1 \leq 2k \leq 2K$ is partly due to the fact that for any $p \in \mathbb{N}$, the tensor $M^{2p}$ can be computed from the first $p$ homogenized tensors $(\mathcal{X}^i(y))_{0\leq i\leq p}$ only (item 4 of Proposition 10.3.8).

Since we have verified that all the requirements of Proposition 10.4.1 are satisfied with $v_\epsilon^* \equiv v_{\epsilon,K}^*$ and $L \equiv 2K+1$ (recall that $M^{2K+1} = 0$ by Corollary 10.3.1), we are in position to state our main result.

**Corollary 10.4.2.** *The error estimate* (10.14) *holds for the reconstructed solution* $W_{\epsilon,2K+1}(v_{\epsilon,K}^*)$*, where* $v_{\epsilon,K}^*$ *is the solution of* (10.59).

## 10.5 Retrieving the classical regimes: low volume fraction limits when the size of the obstacles tends to zero

The goal of this section is to show that our higher-order homogenized models (10.6) have the potential of being valid in any regime of size of holes. For this purpose, we obtain asymptotics for the tensors $\mathcal{X}^{k*}$ and $M^k$ in the low volume fraction limit when the scaling $\eta$ of the obstacle vanishes to zero, $\eta \to 0$. Our main results are stated in Corollaries 10.5.1 and 10.5.2; they imply that both the infinite-order homogenized equation (10.40) and the effective model (10.6) of order $2K+2$ converge coefficientwise to either of the three classical regimes of the literature (namely, to the original Laplace equation (10.1) or to the analogue of the Brinkman or Darcy equation ((10.3) and (10.4)) if $K = 0$, and to either of equations (10.3) or (10.4) for $K \geq 1$ if $\eta$ remains greater than or comparable to the critical size $\eta_{\text{crit}} \sim \eta^{2/(d-2)}$.

In this subsection, we assume, for simplicity, that the space dimension is greater than 3,

$$d \geq 3.$$

We do not consider the case $d = 2$, which requires a specific treatment, although very similar results could be stated (see, e. g., [4]).

In what follows, the hole $\eta T$ is assumed to be strictly included in the unit cell for any $\eta \leq 1$ (it does not touch the boundary): $\eta T \subset\subset P$. Functions of the rescaled cell $\eta^{-1}P$ are indicated by a tilde ˜notation. For a given function $\tilde{v} \in L^2(\eta^{-1}P)$, we denote by $\langle\tilde{v}\rangle$ the average $\langle\tilde{v}\rangle := \eta^d \int_{\eta^{-1}P} \tilde{v}(y)\mathrm{d}y$. With a small abuse of notation, when $\tilde{v} \in L^2(\eta^{-1}P\backslash T)$, we still denote by $\langle\tilde{v}\rangle$ this quantity where we implicitly extend $\tilde{v}$ by 0 within $T$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Let us recall that for any $v \in H^1(P \setminus (\eta T))$, if $\widetilde{v}$ is the rescaled function defined by $\widetilde{v}(y) := v(\eta y)$ in the rescaled cell $\eta^{-1}P \setminus T$, then the $L^2$ norms of $v$ and $\widetilde{v}$ and their gradients are related by the following identities:

$$\|v\|_{L^2(P\setminus(\eta T))} = \eta^{d/2}\|\widetilde{v}\|_{L^2(\eta^{-1}P\setminus T)},$$
$$\|\nabla v\|_{L^2(P\setminus(\eta T),\mathbb{R}^d)} = \eta^{d/2-1}\|\nabla\widetilde{v}\|_{L^2(\eta^{-1}P\setminus T,\mathbb{R}^d)}.$$

We follow the methodology of [4] (and also [37]), which relies extensively on the use of the next two lemmas.

**Lemma 10.5.1.** *Let $d \geq 3$. There exists a constant $C > 0$ independent of $\eta > 0$ such that for any $\widetilde{v} \in H^1(\eta^{-1}P \setminus T)$ that vanishes on the hole $\partial T$ and that is $\eta^{-1}P$-periodic, we have the following inequalities:*

$$\|\widetilde{v}\|_{L^2(\eta^{-1}P\setminus T)} \leq C\eta^{-d/2}\|\nabla\widetilde{v}\|_{L^2(\eta^{-1}P\setminus T,\mathbb{R}^d)}, \tag{10.74}$$
$$\left|\langle\widetilde{v}\rangle\right| \leq C\|\nabla\widetilde{v}\|_{L^2(\eta^{-1}P\setminus T,\mathbb{R}^d)}, \tag{10.75}$$
$$\left\|\widetilde{v} - \langle\widetilde{v}\rangle\right\|_{L^2(\eta^{-1}P\setminus T)} \leq C\eta^{-1}\|\nabla\widetilde{v}\|_{L^2(\eta^{-1}P\setminus T,\mathbb{R}^d)}, \tag{10.76}$$
$$\left\|\widetilde{v} - \langle\widetilde{v}\rangle\right\|_{L^{2d/(d-2)}(\eta^{-1}P\setminus T)} \leq C\|\nabla\widetilde{v}\|_{L^2(\eta^{-1}P\setminus T,\mathbb{R}^d)}. \tag{10.77}$$

*Proof.* See [4, 38]. □

**Lemma 10.5.2.** *Consider $\widetilde{h} \in L^2(\eta^{-1}P \setminus T)$ and let $\widetilde{v} \in H^1(\eta^{-1}P \setminus T)$ be the unique solution to the Poisson problem*

$$\begin{cases} -\Delta\widetilde{v} = \widetilde{h} \quad \text{in } \eta^{-1}P \setminus T, \\ \widetilde{v} = 0 \quad \text{on } \partial T, \\ \widetilde{v} \text{ is } \eta^{-1}P \setminus T\text{-periodic.} \end{cases} \tag{10.78}$$

*There exists a constant $C > 0$ independent of $\eta$ and $\widetilde{h}$ such that*

$$\|\nabla\widetilde{v}\|_{L^2(\eta^{-1}P\setminus T,\mathbb{R}^d)} \leq C(\eta^{-1}\|\widetilde{h} - \langle\widetilde{h}\rangle\|_{L^2(\eta^{-1}P\setminus T)} + \eta^{-d}|\langle\widetilde{h}\rangle|).$$

*Proof.* An integration by parts and Lemma 10.5.1 yield

$$\begin{aligned}
\|\nabla\widetilde{v}\|_{L^2(\eta^{-1}P\setminus T,\mathbb{R}^d)} &= \int_{\eta^{-1}P\setminus T} \widetilde{h}\widetilde{v}\,\mathrm{d}y \\
&= \int_{\eta^{-1}P\setminus T} (\widetilde{h} - \langle\widetilde{h}\rangle)(\widetilde{v} - \langle\widetilde{v}\rangle)\,\mathrm{d}y + \int_{\eta^{-1}P\setminus T} \langle\widetilde{h}\rangle\langle\widetilde{v}\rangle\,\mathrm{d}y \\
&\leq C(\|\widetilde{h} - \langle\widetilde{h}\rangle\|_{L^2(\eta^{-1}P\setminus T)}\|\widetilde{v} - \langle\widetilde{v}\rangle\|_{L^2(\eta^{-1}P\setminus T)} + \eta^{-d}|\langle\widetilde{h}\rangle|\,|\langle\widetilde{v}\rangle|) \\
&\leq C(\eta^{-1}\|\widetilde{h} - \langle\widetilde{h}\rangle\|_{L^2(\eta^{-1}P\setminus T)} + \eta^{-d}|\langle\widetilde{h}\rangle|)\|\nabla\widetilde{v}\|_{L^2(\eta^{-1}P\setminus T,\mathbb{R}^d)}.
\end{aligned}$$

□

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

We also need the so-called Deny–Lions (or Beppo Levi) space $\mathcal{D}^{1,2}(\mathbb{R}^d \setminus T)$ (the reader is referred to [3–5] and also to [44, p. 59] for more detail).

**Definition 10.5.1** (Deny–Lions space). The Deny–Lions space $\mathcal{D}^{1,2}(\mathbb{R}^d \setminus T)$ is the completion of the space of smooth functions by the $L^2$ norm of their gradients:

$$\mathcal{D}^{1,2}(\mathbb{R}^d \setminus T) := \overline{\mathcal{D}(\mathbb{R}^d \setminus T)}^{\|\nabla \cdot\|_{L^2(\mathbb{R}^d \setminus T, \mathbb{R}^d)}}.$$

When $d \geq 3$, it is admits the following characterization:

$$\begin{aligned}&\mathcal{D}^{1,2}(\mathbb{R}^d \setminus T)\\ &\quad = \{\phi \text{ measurable} \mid \|\phi\|_{L^{2d/(d-2)}(\mathbb{R}^d \setminus T)} < +\infty \text{ and } \|\nabla \phi\|_{L^2(\mathbb{R}^d \setminus T, \mathbb{R}^d)} < +\infty\}.\end{aligned}$$

We introduce the unique solution $\Psi$ to the exterior problem

$$\begin{cases} -\Delta \Psi = 0 & \text{in } \mathbb{R}^d \setminus T, \\ \Psi = 0 & \text{on } \partial T, \\ \Psi \to 1 & \text{at } \infty, \end{cases} \tag{10.79}$$

and we denote by $F$ the normal flux

$$F := \int_{\mathbb{R}^d \setminus T} |\nabla \Psi|^2 \mathrm{d}x = -\int_{\partial T} \nabla \Psi \cdot \boldsymbol{n} \mathrm{d}s, \tag{10.80}$$

where $\boldsymbol{n}$ is the normal pointing *inward* $T$. The condition $\Psi \to 1$ at $\infty$ is to be understood in the sense that $\Psi - 1 \in \mathcal{D}^{1,2}(\mathbb{R}^d \setminus T)$.

The following result provides asymptotics for the tensors $\mathcal{X}^k$ and their averages $\mathcal{X}^{k*}$. It extends Theorem 3.1 of [4] (see also [37]), where the particular case $k = 0$ was obtained for the Stokes system.

**Proposition 10.5.1.** *Let $d \geq 3$. For any $k \geq 0$, denote by $\widetilde{\mathcal{X}}^{2k}$ and $\widetilde{\mathcal{X}}^{2k+1}$ the rescaled tensors for any $x \in \eta^{-1} P \setminus T$ by*

$$\widetilde{\mathcal{X}}^{2k}(x) := \eta^{(d-2)(k+1)} \mathcal{X}^{2k}(\eta x) \quad \text{and} \quad \widetilde{\mathcal{X}}^{2k+1}(x) := \eta^{(d-2)(k+1)} \mathcal{X}^{2k+1}(\eta x).$$

*Then:*

1. *there exists a constant $C > 0$ independent of $\eta > 0$ such that*

$$\forall \eta > 0, \quad \left\|\nabla \widetilde{\mathcal{X}}^{2k}\right\|_{L^2(\eta^{-1} P \setminus T, \mathbb{R}^d)} \leq C \quad \text{and} \quad \left\|\nabla \widetilde{\mathcal{X}}^{2k+1}\right\|_{L^2(\eta^{-1} P \setminus T, \mathbb{R}^d)} \leq C; \tag{10.81}$$

2. *the following convergences hold as $\eta \to 0$:*

$$\widetilde{\mathcal{X}}^{2k} \rightharpoonup \frac{\Psi}{F^{k+1}} J^{2k} \quad \text{weakly in } H^1_{\mathrm{loc}}(\mathbb{R}^d \setminus T), \tag{10.82}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$$\widetilde{\mathcal{X}}^{2k+1} \rightharpoonup 0 \quad \textit{weakly in } H^1_{\mathrm{loc}}(\mathbb{R}^d \setminus T), \tag{10.83}$$

$$\mathcal{X}^{2k*} \sim \frac{1}{\eta^{(d-2)(k+1)} F^{k+1}} J^{2k}, \tag{10.84}$$

*where we recall* $J^{2k} := \overbrace{I \otimes I \otimes \cdots \otimes I}^{k \text{ times}}$ *(definition* (10.18)*).*

**Remark 10.5.1.** Let us recall that we already know that $\mathcal{X}^{2k+1*} = 0$ for any $k \in \mathbb{N}$ (Proposition 10.3.3).

*Proof.* The result is proved by induction.

1. *Case* $2k$ *with* $k = 0$. The tensor $\widetilde{\mathcal{X}}^0$ satisfies

$$-\Delta \widetilde{\mathcal{X}}^0 = \eta^d \quad \text{in } \eta^{-1}P \setminus T \tag{10.85}$$

and the other boundary conditions (10.78). Hence Lemma 10.5.2 yields

$$\left\| \nabla \widetilde{\mathcal{X}}^0 \right\|_{L^2(\eta^{-1}P \setminus T, \mathbb{R}^d)} \leq C \eta^{-d} \eta^d |\langle 1 \rangle| \leq C.$$

By (10.75) the average $\langle \widetilde{\mathcal{X}}^0 \rangle$ is bounded. Therefore there exist a constant $c^0 \in \mathbb{R}$ and a function $\hat{\Psi}^0$ such that, up to extracting a subsequence,

$$\left\langle \widetilde{\mathcal{X}}^0 \right\rangle \to c^0 \quad \text{and} \quad \widetilde{\mathcal{X}}^0 \rightharpoonup \hat{\Psi}^0 \quad \text{weakly in } H^1_{\mathrm{loc}}(\mathbb{R}^d \setminus T) \text{ as } \eta \to 0.$$

Furthermore, the lower semicontinuity of the $H^1_{\mathrm{loc}}(\mathbb{R}^d \setminus T)$ norm and (10.77) imply that $\hat{\Psi}^0 - c^0$ belongs to $\mathcal{D}^{1,2}(\mathbb{R}^d \setminus T)$ (see [4] for a detailed justification). Multiplying (10.85) by a compactly supported test function $\Phi \in \mathcal{C}^\infty_c(\mathbb{R}^d \setminus \overline{T})$ and integrating by part yield

$$\int_{\eta^{-1}P \setminus T} \nabla \widetilde{\mathcal{X}}^0 \cdot \nabla \Phi \mathrm{d}y = \eta^d \int_{\eta^{-1}P \setminus T} \Phi \mathrm{d}y.$$

Passing to the limit as $\eta \to 0$ entails that $\hat{\Psi}^0$ is solution to the exterior problem

$$\begin{cases} -\Delta \hat{\Psi}^0 = 0 & \text{in } \mathbb{R}^d \setminus T, \\ \hat{\Psi}^0 = 0 & \text{on } \partial T, \\ \hat{\Psi}^0 \to c^0 & \text{at } \infty. \end{cases} \tag{10.86}$$

By linearity the unique solution to this problem is given by $\widehat{\Psi} = c^0 \Psi$, where $\Psi$ is the solution to equation (10.79). Finally, the constant $c^0$ can be identified by integrating equations (10.85) against the constant test function $\Phi = 1$, which implies

$$c^0 F = -\int_{\partial T} \nabla (c^0 \Psi^0) \cdot \boldsymbol{n} \mathrm{d}s = \lim_{\eta \to 0} -\int_{\partial T} \nabla \widetilde{\mathcal{X}}^0 \cdot \boldsymbol{n} \mathrm{d}s = 1.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Therefore $c^0 = 1/F$, from which (10.82) follows for $k = 0$. Since the obtained limit is unique, the convergence holds for the whole sequence. Then (10.84) follows from a simple change of variable.

2. *Case* $2k+1$ *with* $k = 0$. A simple computation yields

$$-\Delta\widetilde{\mathcal{X}}^1 = 2\eta\partial_j\widetilde{\mathcal{X}}^0 \otimes e_j. \tag{10.87}$$

Applying Lemma 10.5.2 and noting that $\langle 2\eta\partial_j\widetilde{\mathcal{X}}^0\rangle = 0$, we obtain

$$\left\|\nabla\widetilde{\mathcal{X}}^1\right\|_{L^2(\eta^{-1}P\setminus T,\mathbb{R}^d)} \leq C\eta^{-1}\eta\left\|\nabla\widetilde{\mathcal{X}}^0\right\| \leq C'.$$

Integrating equation (10.87) against a compactly supported test function $\Phi \in \mathcal{C}_c^\infty(\mathbb{R}^d \setminus T)$ and passing to the limit as $\eta \to 0$, we obtain with similar arguments the existence of a constant tensor $c^1$ (of order 1) such that, up to the extraction of a subsequence,

$$\widetilde{\mathcal{X}}^1 \rightharpoonup c^1\Psi \quad \text{weakly in } H^1_{\mathrm{loc}}(\mathbb{R}^d \setminus T) \quad \text{and} \quad \langle\widetilde{\mathcal{X}}^1\rangle \to c^1.$$

By Proposition 10.3.3 we have $\langle\widetilde{\mathcal{X}}^1\rangle = 0$, which implies $c^1 = 0$ and hence (10.83) for $k = 0$.

3. *General case*. We now complete the proof by induction on $k$. Assuming that the result holds till rank $k \geq 0$, we compute

$$\begin{aligned}
-\Delta\widetilde{\mathcal{X}}^{2k+2} &= 2\eta^{d-1}\partial_j\widetilde{\mathcal{X}}^{2k+1} \otimes e_j + \eta^d\widetilde{\mathcal{X}}^{2k} \otimes I,\\
-\Delta\widetilde{\mathcal{X}}^{2k+3} &= 2\eta\partial_j\widetilde{\mathcal{X}}^{2k+2} \otimes e_j + \eta^d\widetilde{\mathcal{X}}^{2k+1} \otimes I.
\end{aligned}$$

Applying Lemmas 10.5.2 and 10.5.1 and (10.81) at rank $k$, we obtain

$$\begin{aligned}
&\left\|\nabla\widetilde{\mathcal{X}}^{2k+2}\right\|_{L^2(\eta^{-1}P\setminus T,\mathbb{R}^d)}\\
&\quad\leq C\Big(\eta^{d-2}\left\|\nabla\widetilde{\mathcal{X}}^{2k+1}\right\|_{L^2(\eta^{-1}P\setminus T,\mathbb{R}^d)} + \eta^{d-1}\left\|\widetilde{\mathcal{X}}^{2k} - \langle\widetilde{\mathcal{X}}^{2k}\rangle\right\|_{L^2(\eta^{-1}P\setminus T)}\\
&\qquad + \left\|\langle\widetilde{\mathcal{X}}^{2k}\rangle\right\|_{L^2(\eta^{-1}P\setminus T)}\Big)\\
&\quad\leq C\Big(\eta^{d-2}\left\|\nabla\widetilde{\mathcal{X}}^{2k+1}\right\|_{L^2(\eta^{-1}P\setminus T,\mathbb{R}^d)} + \eta^{d-2}\left\|\nabla\widetilde{\mathcal{X}}^{2k}\right\|_{L^2(\eta^{-1}P\setminus T,\mathbb{R}^d)}\\
&\qquad + \left\|\nabla\widetilde{\mathcal{X}}^{2k}\right\|_{L^2(\eta^{-1}P\setminus T,\mathbb{R}^d)}\Big) \leq C',
\end{aligned} \tag{10.88}$$

$$\begin{aligned}
&\left\|\nabla\widetilde{\mathcal{X}}^{2k+3}\right\|^2_{L^2(\eta^{-1}P\setminus T,\mathbb{R}^d)}\\
&\quad\leq C\Big(\left\|\nabla\widetilde{\mathcal{X}}^{2k+2}\right\|_{L^2(\eta^{-1}P\setminus T,\mathbb{R}^d)} + \eta^{d-1}\left\|\widetilde{\mathcal{X}}^{2k+1} - \langle\widetilde{\mathcal{X}}^{2k+1}\rangle\right\|_{L^2(\eta^{-1}P\setminus T)}\\
&\qquad + \left|\langle\widetilde{\mathcal{X}}^{2k+1}\rangle\right|\Big)\\
&\quad\leq C\Big(\left\|\nabla\widetilde{\mathcal{X}}^{2k+2}\right\|_{L^2(\eta^{-1}P\setminus T,\mathbb{R}^d)} + \eta^{d-2}\left\|\nabla\widetilde{\mathcal{X}}^{2k+1}\right\|\Big) \leq C'.
\end{aligned}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

This implies (10.81) at rank $k+1$. Using similar arguments as previously, we infer the existence of constant tensors $c^{2k+2}$ and $c^{2k+3}$ such that

$$\widetilde{\mathcal{X}}^{2k+2} \rightharpoonup c^{2k+2}\Psi \quad \text{weakly in } H^1_{\text{loc}}(\mathbb{R}^d \setminus T) \quad \text{and} \quad \langle \widetilde{\mathcal{X}}^{2k+2} \rangle \to c^{2k+2},$$
$$\widetilde{\mathcal{X}}^{2k+3} \rightharpoonup c^{2k+3}\Psi \quad \text{weakly in } H^1_{\text{loc}}(\mathbb{R}^d \setminus T) \quad \text{and} \quad \langle \widetilde{\mathcal{X}}^{2k+3} \rangle \to c^{2k+3}.$$

Since we know from Proposition 10.3.3 that $\langle \widetilde{\mathcal{X}}^{2k+3} \rangle = 0$, we obtain $c^{2k+3} = 0$, which yields (10.83) at rank $k+1$. Finally, we integrate (10.88) by parts against the test function $\Phi = 1$ to identify $c^{2k+2}$:

$$\begin{aligned} c^{2k+2}F &= -\int_{\partial T} \nabla\big(c^{2k+2}\Psi\big) \cdot \boldsymbol{n} \mathrm{d}s = \lim_{\eta\to 0} -\int_{\partial T} \nabla \widetilde{\mathcal{X}}^{2k+2} \cdot \boldsymbol{n} \mathrm{d}s \\ &= \lim_{\eta\to 0} \big\langle \widetilde{\mathcal{X}}^{2k} \big\rangle \otimes I = c^{2k} \otimes I. \end{aligned}$$

This implies $c^{2k+2} = J^{2k+2}/F^{k+2}$ and $c^{2k+3} = 0$, which concludes the proof. □

**Remark 10.5.2.** The convergence (10.83) seems to indicate that we may have not found the optimal scaling for the odd-order tensors $\mathcal{X}^{2k+1}$ since we are able to identify only a zero weak limit in (10.83). A more elaborate analysis could be carried on by different techniques; see, e. g., the use of layer potentials in [37].

We are now able to identify the asymptotic behavior of the constant tensors $M^k$. Recall that we already know that $M^{2k+1} = 0$ from Corollary 10.3.1.

**Corollary 10.5.1.** *Let $d \geq 3$. The following convergences hold for the tensors $M^k$ as $\eta \to 0$:*

$$M^0 \sim \eta^{d-2} F, \tag{10.89}$$
$$M^2 \to -I, \tag{10.90}$$
$$\forall k > 1, \quad M^{2k} = o\bigg(\frac{1}{\eta^{(d-2)(k-1)}}\bigg). \tag{10.91}$$

*Proof.* We replace the asymptotics of Proposition 10.5.1 in the explicit formula (10.41) for the tensor $M^k$. Relation (10.89) is a consequence of the definition $M^0 = (\mathcal{X}^{0*})^{-1}$. Convergence (10.90) is obtained by writing

$$M^2 = -\big((\mathcal{X}^{0*})^{-1}\big)^2 \mathcal{X}^{2*} \sim -\frac{\eta^{2(d-2)}F^2}{\eta^{2(d-2)}F^2} I = -\mathrm{I}.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Let us now prove (10.91). By eliminating terms of odd orders in (10.41), we may write, for any $k \geq 1$,

$$\begin{aligned} M^{2k} &= \sum_{p=1}^{2k} \frac{(-1)^p}{(\mathcal{X}^{0*})^{p+1}} \sum_{\substack{i_1+\cdots+i_p=k \\ 1\leq i_1 \ldots i_p \leq k}} \mathcal{X}^{2i_1 *} \otimes \cdots \otimes \mathcal{X}^{2i_p *} \\ &= \sum_{p=1}^{2k} (-1)^p \eta^{(p+1)(d-2)} F^{p+1} \sum_{\substack{i_1+\cdots+i_p=k \\ 1\leq i_1 \ldots i_p \leq k}} \frac{J^{2i_1}}{\eta^{(d-2)(i_1+1)} F^{i_1+1}} \otimes \cdots \otimes \frac{J^{2i_p}}{\eta^{(d-2)(i_p+1)} F^{i_p+1}} \\ &\quad + o\left( \frac{1}{\eta^{(k-1)(d-2)}} \right) \\ &= \frac{J^{2k}}{\eta^{(k-1)(d-2)} F^{k-1}} \left( \sum_{p=1}^{2k} (-1)^p \sum_{\substack{i_1+\cdots+i_p=k \\ 1\leq i_1 \ldots i_p \leq k}} 1 \right) + o\left( \frac{1}{\eta^{(k-1)(d-2)}} \right). \end{aligned}$$

Then (10.91) results from the last summation being zero:

$$\forall k > 1, \quad \sum_{p=1}^{k} (-1)^p \sum_{\substack{i_1+\cdots+i_p=k \\ 1\leq i_1 \ldots i_p \leq k}} 1 = 0. \tag{10.92}$$

There are several ways to obtain the latter formula. A rather direct argument in the spirit of the proof of Proposition 10.3.6 is to apply identity (10.42) to the power series $1/(1-z) = \sum_{k\in\mathbb{N}} z^k$, which yields

$$1 - z = 1 + \sum_{k=1}^{+\infty} \left( \sum_{p=1}^{k} (-1)^p \sum_{\substack{i_1+\cdots+i_p=k \\ 1\leq i_1 \ldots i_p \leq k}} 1 \right) z^k,$$

from which (10.92) follows by identifying the powers in $z^k$. □

**Remark 10.5.3.** We retrieve formally in Corollary 10.5.1 the different classical asymptotic regimes (10.3)–(10.5) for the perforated problem (10.1) (at least for $K = 0$), because the coefficients of (10.9) read as $\eta \to 0$:

$$\begin{aligned} \epsilon^{-2} M^0 &\sim \frac{\eta^{d-2}}{\epsilon^2} F, \\ \epsilon^0 M^2 &\to -I, \\ \epsilon^{2k-2} M^{2k} &= o\left( \left( \frac{\epsilon^2}{\eta^{d-2}} \right)^{k-1} \right) \quad \text{for } k \geq 1. \end{aligned}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

These asymptotics bring into play the ratio $\epsilon^2/\eta^{d-2}$, and so the critical scaling $\eta \sim \epsilon^{2/(d-2)}$ corresponding to the "Brinkman" regime (10.3) (which implies $\epsilon^{-2}M^0 \to F$ and $\epsilon^{2k+2}M^{2k} = o(1)$). The Darcy regimes (10.4) and (10.5) correspond to the situation where $\eta^{d-2}/\epsilon^2 \to +\infty$; in that case the zeroth-order term $\epsilon^{-2}M^0$ is dominant. The Stokes regime (10.2) is found for $K = 0$ and $\eta = o(\epsilon^{2/(d-2)})$ (which implies $\epsilon^{-2}M^0 \to 0$). Note that $\epsilon^0 M^2 \to -I$ whatever the scaling $\eta \to 0$. Unfortunately, the inverse of the critical ratio $\epsilon^2/\eta^{d-2}$ can blow up as $\eta$ vanishes at the rate $\eta = o(\epsilon^{2/(d-2)})$: a more elaborate analysis will be performed in future works to determine whether $\epsilon^{2k-2}M^{2k}$ with $k > 1$ still converges to zero in this regime.

We conclude this paper with the statement that the regimes (10.4) and (10.5) are also captured in the low volume fraction limit as $\eta \to 0$ by the homogenized equation (10.6) of finite order $2K + 2$. Because of Proposition 10.4.4, it suffices to establish that $\mathbb{D}_K^{2K+2}$ satisfies the same asymptotics as $M^{2K+2}$ (hence all the coefficients $\mathbb{D}_K^{2k}$ for $0 \le 2k \le 2K + 2$ because of (10.4.4)). The proof of this result requires the following asymptotic bounds for the tensors $N^k$.

**Proposition 10.5.2.** *Let $d \ge 3$. For any $k \ge 0$, let the rescaled tensors $\widetilde{N}^{2k}$ and $\widetilde{N}^{2k+1}$ in $\eta^{-1}P \setminus T$ be defined by*

$$\forall y \in \eta^{-1}P \setminus T, \quad \widetilde{N}^{2k}(y) := \eta^{(d-2)k} N^{2k}(\eta y) \quad \text{and} \quad \widetilde{N}^{2k+1}(y) := \eta^{(d-2)k} N^{2k+1}(\eta y).$$

*Then there exists a constant $C$ independent of $\eta > 0$ such that*

$$\forall \eta > 0, \quad \left\|\nabla \widetilde{N}^{2k}\right\|_{L^2(\eta^{-1}P\setminus T)} \le C \quad \text{and} \quad \left\|\nabla \widetilde{N}^{2k+1}\right\|_{L^2(\eta^{-1}P\setminus T)} \le C. \tag{10.93}$$

*Moreover, the following convergences hold as $\eta \to 0$:*

$$\widetilde{N}^0 \rightharpoonup \Psi \quad \text{weakly in } H^1_{\mathrm{loc}}(\mathbb{R}^d \setminus T), \tag{10.94}$$

$$\forall k \ge 1, \quad \widetilde{N}^k \rightharpoonup 0 \quad \text{weakly in } H^1_{\mathrm{loc}}(\mathbb{R}^d \setminus T). \tag{10.95}$$

*Proof.* Using definition (10.44) of the tensors $N^k$ and eliminating odd-order terms, the tensors $\widetilde{N}^{2k}$ and $\widetilde{N}^{2k+1}$ can be rewritten as

$$\widetilde{N}^0 = \frac{M^0}{\eta^{d-2}} \widetilde{\mathcal{X}}^0,$$

$$\forall k \ge 1, \quad \widetilde{N}^{2k} = \frac{M^0}{\eta^{d-2}} \widetilde{\mathcal{X}}^{2k} + M^2 \otimes \widetilde{\mathcal{X}}^{2k-2} + \sum_{p=2}^{k} \eta^{(d-2)(p-1)} M^{2p} \otimes \widetilde{\mathcal{X}}^{2(k-p)},$$

$$\forall k \ge 0, \quad \widetilde{N}^{2k+1} = \frac{M^0}{\eta^{d-2}} \widetilde{\mathcal{X}}^{2k+1} + M^2 \otimes \widetilde{\mathcal{X}}^{2k-1} + \sum_{p=2}^{k} \eta^{(d-2)(p-1)} M^{2p} \otimes \widetilde{\mathcal{X}}^{2(k-p)+1}.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

By Proposition 10.5.1 and Corollary 10.5.1 we obtain

$$\begin{aligned}
\widetilde{N}^0 &= F\widetilde{\mathcal{X}}^0 + o_{\mathcal{D}^{1,2}(\mathbb{R}^d\setminus T)}(1),\\
\forall k \geq 1, \quad \widetilde{N}^{2k} &= F\widetilde{\mathcal{X}}^{2k} - \widetilde{\mathcal{X}}^{2k-2}\otimes I + o_{\mathcal{D}^{1,2}(\mathbb{R}^d\setminus T)}(1),\\
\forall k \geq 1, \quad \widetilde{N}^{2k+1} &= F\widetilde{\mathcal{X}}^{2k+1} - \widetilde{\mathcal{X}}^{2k-1}\otimes I + o_{\mathcal{D}^{1,2}(\mathbb{R}^d\setminus T)}(1),
\end{aligned}$$

where $o_{\mathcal{D}^{1,2}(\mathbb{R}^d\setminus T)} \to 0$ in $\mathcal{D}^{1,2}(\mathbb{R}^d \setminus T)$. This implies (10.93)–(10.95) by using once again Proposition 10.5.1. □

**Corollary 10.5.2.** *As $\eta \to 0$, we have:*

$$\begin{aligned}
\mathbb{D}_0^2 &\to -I,\\
\mathbb{D}_K^{2K+2} &= o\left(\frac{1}{\eta^{(d-2)K}}\right) = o\left(\frac{1}{\eta^{(d-2)((K+1)-1)}}\right) \quad \textit{for any } K \geq 1.
\end{aligned}$$

*Proof.* According to (10.69), $\mathbb{D}_0^2 = -\int_Y N^0\otimes N^0\otimes I\mathrm{d}y$. Then Proposition 10.5.1 and Corollary 10.5.1 imply

$$\begin{aligned}
-\int_Y N^0\otimes N^0\otimes I\mathrm{d}y &= -(M^0)^2\int_Y |\mathcal{X}^0|^2\mathrm{d}yI = -(M^0)^2\eta^{-2(d-2)}\eta^d\int_Y |\widetilde{\mathcal{X}}^0|^2\mathrm{d}yI\\
&\sim -F^2\langle\widetilde{\mathcal{X}}^0\rangle^2 I \to -I.
\end{aligned}$$

Finally, for $K \geq 1$, we estimate $\mathbb{D}_K^{2K+2}$ as follows:

$$\begin{aligned}
\mathbb{D}_K^{2K+2} &= (-1)^K\eta^d\eta^{-(d-2)2\lfloor K/2\rfloor}\int_{\eta^{-1}P\setminus T}(\widetilde{N}^K - \langle\widetilde{N}^K\rangle)\otimes(\widetilde{N}^K - \langle\widetilde{N}^K\rangle)\mathrm{d}y\\
&= O(\eta^{-(d-2)(2\lfloor K/2\rfloor-1)}) = o(\eta^{-(d-2)K}).
\end{aligned}$$

□

# Bibliography


[1] A. Abdulle and T. Pouchon. Effective models for the multidimensional wave equation in heterogeneous media over long time and numerical homogenization. *Math. Models Methods Appl. Sci.*, 26(14):2651–2684, 2016.

[2] A. Abdulle and T. Pouchon. Effective models and numerical homogenization for wave propagation in heterogeneous media on arbitrary timescales. *arXiv preprint* arXiv:1905.09062, 2019.

[3] G. Allaire. Homogénéisation des équations de Stokes et de Navier–Stokes. PhD thesis, Université Paris 6, 1989.

[4] G. Allaire. Continuity of the Darcy's law in the low-volume fraction limit. *Ann. Sc. Norm. Super. Pisa, Cl. Sci. (4)*, 18(4):475–499, 1991.

[5] G. Allaire. Homogenization of the Navier–Stokes equations with a slip boundary condition. *Commun. Pure Appl. Math.*, 44(6):605–641, 1991.


EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[6] G. Allaire. Homogenization and two-scale convergence. *SIAM J. Math. Anal.*, 23(6):1482–1518, 1992.
[7] G. Allaire. *Shape Optimization by the Homogenization Method*, volume 146. Springer Science & Business Media, 2012.
[8] G. Allaire and M. Amar. Boundary layer tails in periodic homogenization. *ESAIM Control Optim. Calc. Var.*, 4:209–243, 1999.
[9] G. Allaire, M. Briane, and M. Vanninathan. A comparison between two-scale asymptotic expansions and Bloch wave expansions for the homogenization of periodic structures. *SeMA J.*, 73(3):237–259, 2016.
[10] G. Allaire and C. Conca. Bloch wave homogenization and spectral asymptotic analysis. *J. Math. Pures Appl.*, 77(2):153–208, 1998.
[11] G. Allaire, P. Geoffroy-Donders, and O. Pantz. Topology optimization of modulated and oriented periodic microstructures by the homogenization method. *Comput. Math. Appl.*, 78(7):2197–2229, 2019.
[12] G. Allaire, A. Lamacz, and J. Rauch. Crime pays; homogenized wave equations for long times. *arXiv preprint* arXiv:1803.09455, 2018.
[13] G. Allaire and T. Yamada. Optimization of dispersive coefficients in the homogenization of the wave equation in periodic structures. *Numer. Math.*, 140(2):265–326, 2018.
[14] I. Babuška. Homogenization and its application. Mathematical and computational problems. In *Numerical Solution of Partial Differential Equations – III*, pages 89–116. Elsevier, 1976.
[15] N. Bakhvalov and G. Panasenko. *Homogenisation: Averaging Processes in Periodic Media*, volume 36 of Mathematics and its Applications (Soviet Series). Kluwer Academic Publishers Group, Dordrecht, 1989.
[16] M. P. Bendsøe and N. Kikuchi. Generating optimal topologies in structural design using a homogenization method. *Comput. Methods Appl. Mech. Eng.*, 71(2):197–224, 1988.
[17] M. P. Bendsoe and O. Sigmund. *Topology Optimization: Theory, Methods and Applications*. Springer-Verlag, Berlin, 2003.
[18] A. Bensoussan, J.-L. Lions, and G. Papanicolaou. *Asymptotic analysis for periodic structures*, volume 374. American Mathematical Soc., 2011.
[19] F. Blanc and S. A. Nazarov. Asymptotics of solutions to the Poisson problem in a perforated domain with corners. *J. Math. Pures Appl.*, 76(10):893–911, 1997.
[20] T. Borrvall and J. Petersson. Large-scale topology optimization in 3D using parallel computing. *Comput. Methods Appl. Mech. Eng.*, 190(46–47):6201–6229, 2001.
[21] T. Borrvall and J. Petersson. Topology optimization of fluids in Stokes flow. *Int. J. Numer. Methods Fluids*, 41(1):77–107, 2003.
[22] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer Science & Business Media, 2010.
[23] K. D Cherednichenko and J. A Evans. Full two-scale asymptotic expansion and higher-order constitutive laws in the homogenization of the system of quasi-static Maxwell equations. *Multiscale Model. Simul.*, 14(4):1513–1539, 2016.
[24] A. Cherkaev. *Variational Methods for Structural Optimization*, volume 140. Springer Science & Business Media, 2012.
[25] D. Cioranescu and F. Murat. A strange term coming from nowhere. In *Topics in the Mathematical Modelling of Composite Materials*, volume 31 of Progress in Nonlinear Differential Equations and Their Applications, pages 45–93. Birkhäuser Boston, Boston, MA, 1997.
[26] C. Conca, F. Murat, and O. Pironneau. The Stokes and Navier–Stokes equations with boundary conditions involving the pressure. *Jpn. J. Math. New Ser.*, 20(2):279–318, 1994.
[27] E. M. Dede. Multiphysics topology optimization of heat transfer and fluid flow systems. In *Proceedings of the COMSOL Users Conference*, 2009.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[28] S. B. Dilgen, C. B. Dilgen, D. R. Fuhrman, O. Sigmund, and B. S. Lazarov. Density based topology optimization of turbulent flow heat transfer systems. *Struct. Multidiscip. Optim.*, 57(5):1905–1918, May 2018.
[29] P. Donato and J. S. J. Paulin. Homogenization of the Poisson equation in a porous medium with double periodicity. *Jpn. J. Ind. Appl. Math.*, 10(2):333, 1993.
[30] P. G. Donders. Homogenization method for topology optimization of structures built with lattice materials. Thèse, Université Paris-Saclay, December 2018.
[31] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*, volume 159. Springer Science & Business Media, 2013.
[32] L. C. Evans. *Partial Differential Equations*, volume 19 of Graduate Studies in Mathematics, 2nd edition. American Mathematical Society, Providence, RI, 2010.
[33] F. Feppon. Shape and topology optimization of multiphysics systems. PhD thesis, Thèse de doctorat de l'Université Paris Saclay préparée à l'École polytechnique, 2019.
[34] F. Feppon. High order homogenization of the Stokes system in a periodic porous medium. *SIAM J. Math. Anal.*, 53(3):2890–2924, 2021.
[35] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*, volume 69 of Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011.
[36] J. P Groen and O. Sigmund. Homogenization-based topology optimization for high-resolution manufacturable microstructures. *Int. J. Numer. Methods Eng.*, 113(8):1148–1163, 2018.
[37] W. Jing. A unified homogenization approach for the Dirichlet problem in perforated domains. *arXiv preprint* arXiv:1901.08251, 2019.
[38] O. A Ladyzhenskaya. *The Mathematical Theory of Viscous Incompressible Flow*, volume 2. Gordon and Breach, New York, 1969.
[39] J.-L. Lions. *Perturbations Singulières dans les Problèmes aux Limites et en Contrôle Optimal*, volume 323 of Lecture Notes in Mathematics. Springer-Verlag, Berlin–New York, 1973.
[40] J.-L. Lions. Asymptotic expansions in perforated media with a periodic structure. *Rocky Mt. J. Math.*, 10(1):125–140, 1980.
[41] J.-L. Lions. *Some Methods in the Mathematical Analysis of Systems and Their Control*. Science Press, Beijing, 1981.
[42] W. C. H. McLean. *Strongly Elliptic Systems and Boundary Integral Equations*. Cambridge University Press, 2000.
[43] F. Murat and J. Simon. Sur le contrôle par un domaine géométrique. *Publication du Laboratoire d'Analyse Numérique de l'Université Pierre et Marie Curie*, 1976.
[44] J.-C. Nédélec. *Acoustic and Electromagnetic Equations: Integral Representations for Harmonic Problems*. Springer Science & Business Media, 2001.
[45] O. Pantz and K. Trabelsi. A post-treatment of the homogenization method for shape optimization. *SIAM J. Control Optim.*, 47(3):1380–1398, 2008.
[46] M. Pietropaoli, F. Montomoli, and A. Gaymann. Three-dimensional fluid topology optimization for heat transfer. *Struct. Multidiscip. Optim.*, 59(3):801–812, 2019.
[47] N. Pollini, O. Sigmund, C. S. Andreasen, and J. Alexandersen. A "poor man's" approach for high-resolution three-dimensional topology design for natural convection problems. *Adv. Eng. Softw.*, 140:102736, 2020.
[48] T. N. Pouchon. Effective models and numerical homogenization methods for long time wave propagation in heterogeneous media. PhD thesis, EPFL, 2017.
[49] J. Rauch and M. Taylor. Potential and scattering theory on wildly perturbed domains. *J. Funct. Anal.*, 18:27–59, 1975.
[50] E. Sanchez-Palencia. Fluid flow in porous media. In *Non-Homogeneous Media and Vibration Theory*, pages 129–157, 1980.
[51] F. Santosa and W. W Symes. A dispersive effective medium for wave propagation in periodic composites. *SIAM J. Appl. Math.*, 51(4):984–1005, 1991.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[52] V. P. Smyshlyaev and K. D. Cherednichenko. On rigorous derivation of strain gradient effects in the overall behaviour of periodic heterogeneous media. *J. Mech. Phys. Solids*, 48(6–7):1325–1357, 2000.

[53] L. Tartar. *The General Theory of Homogenization: A Personalized Introduction*, volume 7. Springer Science & Business Media, 2009.

[54] X. Zhao, M. Zhou, O. Sigmund, and C. Andreasen. A "poor man's approach" to topology optimization of cooling channels based on a Darcy flow model. *Int. J. Heat Mass Transf.*, 116:1108–1123, 2018.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Jérôme Lemoine and Arnaud Münch

# 11 Least-squares approaches for the 2D Navier–Stokes system

**Abstract:** We analyze a least-squares approach to approximate weak solutions of the 2D Navier–Stokes system. In the first part, we consider a steady case and introduce a quadratic functional based on a weak norm of the state equation. We construct a minimizing sequence for the functional that strongly converges to a solution of the equation. After a finite number of iterates related to the value of the viscosity constant, the convergence is quadratic, from any initial guess. We then apply iteratively the analysis on the backward Euler scheme associated with the unsteady Navier–Stokes equation and prove the convergence of the iterative process uniformly with respect to the time discretization. In a second part, we reproduce the analysis for the unsteady case by introducing a space-time least-squares functional. This allows us to alleviate the smallness property on the data, assumed the steady case. The method turns out to be related to the globally convergent damped Newton approach applied to the Navier–Stokes operator, in contrast to the standard Newton method used to solve the weak formulation of the Navier–Stokes system. Numerical experiments illustrates our analysis.

**Keywords:** Navier–Stokes equation, implicit time scheme, least-squares approach, space-time variational formulation, damped Newton method

**MSC 2010:** 65L05, 35Q30, 76D05

## 11.1 Introduction. Motivation

Let $\Omega \subset \mathbb{R}^2$ be a bounded connected open set with Lipschitz boundary $\partial\Omega$ is. We denote $\boldsymbol{\mathcal{V}} = \{v \in \mathcal{D}(\Omega)^2, \nabla\cdot v = 0\}$, $\boldsymbol{H}$ the closure of $\boldsymbol{\mathcal{V}}$ in $L^2(\Omega)^2$, and $\boldsymbol{V}$ the closure of $\boldsymbol{\mathcal{V}}$ in $H^1(\Omega)^2$. Endowed with the norm $\|v\|_{\boldsymbol{V}} = \|\nabla v\|_2 := \|\nabla v\|_{(L^2(\Omega))^4}$, $\boldsymbol{V}$ is a Hilbert space. The dual $\boldsymbol{V}'$ of $\boldsymbol{V}$ endowed with the dual norm

$$\|v\|_{\boldsymbol{V}'} = \sup_{w\in\boldsymbol{V},\ \|w\|_{\boldsymbol{V}}=1} \langle v, w\rangle_{\boldsymbol{V}'\times\boldsymbol{V}}$$

**Acknowledgement:** The authors acknowledge the anonymous referees for their very careful reading of this work, which have led to several improvements.

**Jérôme Lemoine, Arnaud Münch,** Laboratoire de Mathématiques Blaise Pascal, Université Clermont Auvergne, UMR CNRS 6620, Campus des Cézeaux, 63177 Aubière, France, e-mails: jerome.lemoine@uca.fr, arnaud.munch@uca.fr

https://doi.org/10.1515/9783110695984-011

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

is also a Hilbert space. We denote by $\langle \cdot, \cdot \rangle_{\boldsymbol{V}'}$ the scalar product associated with the norm $\| \ \|_{\boldsymbol{V}'}$. As usual, the space $\boldsymbol{H}$ is identified with its dual, and the duality $\boldsymbol{V}'$, $\boldsymbol{V}$ is expressed by using $\boldsymbol{H}$ as a pivot space.

Let $T > 0$. We denote $Q_T := \Omega \times (0, T)$ and $\Sigma_T := \partial\Omega \times [0, T]$.

The Navier–Stokes system describes a viscous incompressible fluid flow in the bounded domain $\Omega$ during the time interval $(0, T)$ submitted to the external force $f$. It reads as follows:

$$\begin{cases} y_t - \nu\Delta y + (y \cdot \nabla)y + \nabla p = f, \quad \nabla \cdot y = 0 \quad \text{in } Q_T, \\ y = 0 \quad \text{on } \Sigma_T, \\ y(\cdot, 0) = u_0 \quad \text{in } \Omega, \end{cases} \tag{11.1}$$

where $y$ is the velocity of the fluid, $p$ its pressure, and $\nu$ is the viscosity constant. We refer to [16, 22, 25].

We recall (see [25]) that for $f \in L^2(0, T; H^{-1}(\Omega)^2)$ and $u_0 \in \boldsymbol{H}$, there exists a unique weak solution $y \in L^2(0, T; \boldsymbol{V})$, $\partial_t y \in L^2(0, T; \boldsymbol{V}')$ of the system

$$\begin{cases} \dfrac{d}{dt} \displaystyle\int_\Omega y \cdot w + \nu \int_\Omega \nabla y \cdot \nabla w + \int_\Omega y \cdot \nabla y \cdot w = \langle f, w \rangle_{H^{-1}(\Omega)^2 \times H_0^1(\Omega)^2} \quad \forall w \in \boldsymbol{V}, \\ y(\cdot, 0) = u_0 \quad \text{in } \Omega. \end{cases} \tag{11.2}$$

This work is concerned with the approximation of solution for (11.2), that is, an explicit construction of a sequence $(y_k)_{k \in \mathbb{N}}$ converging to a solution $y$ for a suitable norm. In most of the works devoted to this topic (we refer, for instance, to [7, 19]), the approximation of (11.2) is addressed through a time marching method. Given $\{t_n\}_{n=0,\dots,N}$, $N \in \mathbb{N}$, a uniform discretization of the time interval $(0, T)$ and the corresponding time discretization step $\delta t = T/N$, we mention, for instance, the unconditionally stable backward Euler scheme: for each $n \geq 0$, find a solution $y^{n+1}$ of

$$\begin{cases} \displaystyle\int_\Omega \frac{y^{n+1} - y^n}{\delta t} \cdot w + \nu \int_\Omega \nabla y^{n+1} \cdot \nabla w + \int_\Omega y^{n+1} \cdot \nabla y^{n+1} \cdot w \\ \qquad = \langle f^n, w \rangle_{H^{-1}(\Omega)^2 \times H_0^1(\Omega)^2} \quad \forall n \geq 0, \ \forall w \in \boldsymbol{V}, \\ y^0(\cdot, 0) = u_0 \quad \text{in } \Omega, \end{cases} \tag{11.3}$$

with $f^n := \frac{1}{\delta t} \int_{t_n}^{t_{n+1}} f(\cdot, s) ds$. The piecewise linear interpolation (in time) of $(y^n)_{n \in [0,N]}$ weakly converges in $L^2(0, T; \boldsymbol{V})$ toward a solution $y$ of (11.2) as $\delta t$ goes to zero (we refer to [25, Chapter 3, Section 4]). Moreover, it achieves a first-order convergence with respect to $\delta t$. We also refer to [26] for a stability analysis of the scheme in long time and to [23].

For each $n \geq 0$, the determination of $y^{n+1}$ from $y^n$ requires the resolution of a (nonlinear) steady Navier–Stokes equation, parameterized by $\nu$ and $\delta t$. This can be

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

done using Newton-type methods (see, for instance, [20, Section 10.3] and [5] for the weak formulation of (11.3), which reads as follows: find a solution $y = y^{n+1} \in \boldsymbol{V}$ of

$$\alpha \int_\Omega y \cdot w + \nu \int_\Omega \nabla y \cdot \nabla w + \int_\Omega y \cdot \nabla y \cdot w = \langle f, w \rangle_{H^{-1}(\Omega)^2 \times H_0^1(\Omega)^2} + \alpha \int_\Omega g \cdot w \quad \forall w \in \boldsymbol{V} \tag{11.4}$$

with

$$\alpha := \frac{1}{\delta t} > 0, \quad f := f^n = \frac{1}{\delta t} \int_{t_n}^{t_{n+1}} f(\cdot, s) ds, \quad g = y^n. \tag{11.5}$$

Introducing the application $F : \boldsymbol{V} \times \boldsymbol{V} \to \mathbb{R}$ by

$$\begin{aligned} F(y,z) := & \int_\Omega \alpha y \cdot z + \nu \nabla y \cdot \nabla z + (y \cdot \nabla) y \cdot z \\ & - \langle f, z \rangle_{H^{-1}(\Omega)^2 \times H_0^1(\Omega)^2} - \alpha \int_\Omega g \cdot z = 0 \quad \forall z \in \boldsymbol{V}, \end{aligned} \tag{11.6}$$

the Newton algorithm formally reads

$$\begin{cases} y_0 \in \boldsymbol{V}, \\ D_y F(y_k, w) \cdot (y_{k+1} - y_k) = -F(y_k, w) \quad \forall w \in \boldsymbol{V}, k \geq 0. \end{cases} \tag{11.7}$$

If the initial guess $y_0$ is close enough to a solution of (11.4) (i. e., a solution satisfying $F(y, w) = 0$ for all $w \in \boldsymbol{V}$) and if $D_y F(y_k, \cdot)$ is invertible, then the sequence $(y_k)_k$ converges. We refer to [20, Section 10.3] and [6, Chapter 6]).

Alternatively, we may also employ least-squares methods, which consist in minimizing quadratic functional and which measure how an element $y$ is close to the solution. For instance, we may introduce the extremal problem $\inf_{y \in \boldsymbol{V}} E(y)$ with $E : \boldsymbol{V} \to \mathbb{R}^+$ defined by

$$E(y) := \frac{1}{2} \int_\Omega \alpha |v|^2 + \nu |\nabla v|^2, \tag{11.8}$$

where the corrector $v$ is a unique solution in $\boldsymbol{V}$ of the problem

$$\begin{aligned} \alpha \int_\Omega v \cdot w + \nu \int_\Omega \nabla v \cdot \nabla w = & -\alpha \int_\Omega y \cdot w - \nu \int_\Omega \nabla y \cdot \nabla w - \int_\Omega y \cdot \nabla y \cdot w \\ & + \langle f, w \rangle_{H^{-1}(\Omega)^2 \times H_0^1(\Omega)^2} + \alpha \int_\Omega g \cdot w \quad \forall w \in \boldsymbol{V}. \end{aligned} \tag{11.9}$$

Remark that $E(y) = 0$ is zero if and only if $y \in \boldsymbol{V}$ is a (weak) solution of (11.4), i. e., a zero of $F(y, w) = 0$ for all $w \in \boldsymbol{V}$. As a matter of fact, the infimum is reached.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

The minimization of the functional $E$ over $\boldsymbol{V}$ leads to a "weak" least-squares method. In fact, there is a close connection between $E$ and $F$ through the equality $\sqrt{2E(y)} = \sup_{w\in\boldsymbol{V},w\neq 0} \frac{F(y,w)}{|||w|||_{\boldsymbol{V}}}$, where $|||w|||_{\boldsymbol{V}}$ is defined in (11.11). It follows that $E$ is equivalent to the $\boldsymbol{V}'$-norm of the Navier–Stokes equation (see Remark 11.2.9). The terminology "$H^{-1}$ least-squares method" is employed in [2], where this method has been introduced and numerically implemented to approximate the solutions of (11.2) through the scheme (11.3). We also mention [6, Chapter 4, Section 6], which studied later the use of a least-squares strategy to solve a steady Navier–Stokes equation without incompressibility constraint. Least-squares methods to solve nonlinear boundary value problems have been the subject of intensive developments in the last decades, as they present several advantages, notably on computational and stability viewpoints. We refer to books [1, 8]. In the first part of this work, we rigorously analyze the method introduced in [2, 10] and show that we may construct minimizing sequences in $\boldsymbol{V}$ for $E$ that converge strongly to a solution of (11.2). We then justify the use of that weak least-squares method to solve iteratively scheme (11.3), leading to an approximation of the solution of (11.2). This requires to show some convergence properties of the minimizing sequence for $E$, uniformly with respect to $n$, related to the time discretization. As we will see, this requires smallness assumptions on the data $u_0$ and $f$. In the second part, we extend this analysis to a full space-time setting. More precisely, following the terminology of [2], we introduce the following $L^2(0,T;\boldsymbol{V}')$ least-squares functional $\widetilde{E} : H^1(0,T;\boldsymbol{V}') \cap L^2(0,T;\boldsymbol{V}) \to \mathbb{R}^+$:

$$\widetilde{E}(y) := \frac{1}{2}\left\| y_t + \nu B_1(y) + B(y,y) - f \right\|^2_{L^2(0,T;\boldsymbol{V}')}, \tag{11.10}$$

where $B_1$ and $B$ are defined in Lemmas 11.3.2 and 11.3.3. Again, the real quantity $\widetilde{E}(y)$ measures how the element $y$ is close to the solution of (11.2). The minimization of this functional leads to a so-called continuous weak least-squares type method.

This paper is organized as follows. In Section 11.2, we analyze the least-squares method (11.8)–(11.9) associated with weak solutions of (11.4). We first show that $E$ is differentiable over $\boldsymbol{V}$ and that any critical point for $E$ in a ball $\mathbb{B}$ growing with $\alpha$ is also a zero of $E$. This is done by introducing a descent direction $Y_1$ for $E$ at any point $y \in \boldsymbol{V}$ for which $E'(y)\cdot Y_1$ is proportional to $E(y)$. Then, assuming that there exists a least one solution of (11.4) in $\mathbb{B}$, we show that any minimizing sequence $(y_k)_{(k\in\mathbb{N})}$ for $E$ uniformly in $\mathbb{B}$ strongly converges to a solution of (11.4). Such a limit belongs to $\mathbb{B}$ and is in fact the unique solution. Eventually, we construct a minimizing sequence based on the element $Y_1$ and initialized with $g$ assumed in $\boldsymbol{V}$. We show that if $\alpha$ is large enough, then this particular sequence is uniformly in $\mathbb{B}$ and converges (quadratically after a finite number of iterates related to the values of $\nu$ and $\alpha$) strongly to the solution of (11.4). We also emphasize that this specific sequence coincides with that obtained from the damped Newton method (a globally convergent generalization of (11.7)) and with (11.7) for $\alpha$ large enough. Then, in Section 11.2.4, as an application, we consider the least-

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

squares approach to solve iteratively the backward Euler scheme (11.47). For each $n > 0$, we define a minimizing sequence $(y_k^{n+1})_{k\geq 0}$ based on $Y_1^{n+1}$ to approximate the $y^{n+1}$. Adapting the global convergence result of Section 11.2, we then show, assuming that $\alpha$ is large enough (which is achieved by taking a small enough time discretization step $\delta t$) and a smallness property on $\|u_0\|_2 + \|f\|_{L^2(0,T;(H^{-1}(\Omega))^2)}$, the strong convergence of the minimizing sequences uniformly with respect to the time discretization. The analysis is performed in 2D for weak and regular solutions. In particular, we justify the use of Newton-type methods to solve implicit time schemes for (11.1), as mentioned in [20, Section 10.3]. To the best of our knowledge, such analysis of convergence is original.

In Section 11.3, we reproduce the analysis in a space-time setting with the weak solution of (11.2) associated with initial data $u_0$ in $\boldsymbol{H}$ and source term $f \in L^2(0,T;H^{-1}(\Omega))$. In contrast with the previous section, this setting requires no smallness assumptions on the data.

In Section 11.4, we discuss numerical experiments based on finite element approximations in space for two 2D geometries: the celebrated example of the channel with a backward facing step and the semicircular driven cavity introduced in [9]. We notably exhibit the robustness of the damped Newton method (compared to the Newton one), including for small values of the viscosity constant. Section 11.5 concludes with some perspectives.

This review paper gathers some results for the 2D case obtained in [14] and [13] for the steady and unsteady cases, respectively. These references address the 3D case as well.

## 11.2 Analysis of a least-squares method for a steady Navier–Stokes equation

In this section, we analyze a least-squares method to solve the $\alpha$ steady Navier–Stokes equation (11.4) with $\alpha > 0$: we notably use and adapt some arguments from [15] devoted to the case $\alpha = 0$.

### 11.2.1 Technical preliminary results

We endow the space $\boldsymbol{V}$ with the norm $\|y\|_{\boldsymbol{V}} := \|\nabla y\|_2$ for $y \in \boldsymbol{V}$. We will also use the following notations:

$$|||y|||_{\boldsymbol{V}}^2 := \alpha\|y\|_2^2 + \nu\|\nabla y\|_2^2 \quad \forall y \in \boldsymbol{V} \tag{11.11}$$

and $\langle y,z\rangle_{\boldsymbol{V}} := \alpha\int_\Omega yz + \nu\int_\Omega \nabla y\cdot\nabla z$, so that $\langle y,z\rangle_{\boldsymbol{V}} \leq |||y|||_{\boldsymbol{V}}|||z|||_{\boldsymbol{V}}$ for any $y,z \in \boldsymbol{V}$.

We will repeatedly use the following classical estimates (see [25]).

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Lemma 11.2.1.** *Let $u, v \in \boldsymbol{V}$. Then*

$$-\int_\Omega u \cdot \nabla u \cdot v = \int_\Omega u \cdot \nabla v \cdot u \leq \sqrt{2}\|u\|_2 \|\nabla v\|_2 \|\nabla u\|_2. \tag{11.12}$$

**Definition 11.2.1.** For any $y \in \boldsymbol{V}$, $\alpha > 0$, and $\nu > 0$, we define

$$\tau(y) := \frac{\|y\|_{\boldsymbol{V}}}{\sqrt{2\alpha\nu}}$$

and

$$\mathbb{B} := \{y \in \boldsymbol{V}, \tau(y) < 1\}.$$

We will also repeatedly use the following Young type inequalities.

**Lemma 11.2.2.** *For any $u, v \in \boldsymbol{V}$, we have the following inequality:*

$$\sqrt{2}\|u\|_2 \|\nabla v\|_2 \|\nabla u\|_2 \leq \tau(v)|||u|||_{\boldsymbol{V}}^2. \tag{11.13}$$

Let $f \in H^{-1}(\Omega)^2$, $g \in L^2(\Omega)^2$, and $\alpha \in \mathbb{R}_+^\star$. We have the following:

**Proposition 11.2.3.** *Let $\Omega \subset \mathbb{R}^d$ be bounded and Lipschitz. There exists at least one solution $y$ of* (11.4) *satisfying*

$$|||y|||_{\boldsymbol{V}}^2 \leq \frac{c_0}{\nu}\|f\|_{H^{-1}(\Omega)^d}^2 + \alpha\|g\|_2^2, \tag{11.14}$$

*where $c_0 > 0$, only connected to the Poincaré constant, depends on $\Omega$. If, moreover, $\Omega$ is $C^2$ and $f \in L^2(\Omega)^d$, then any solution $y \in \boldsymbol{V}$ of* (11.4) *belongs to $H^2(\Omega)^2$.*

*Proof.* We refer to [16]. □

**Lemma 11.2.4.** *Let a solution $y \in \boldsymbol{V}$ of* (11.4) *satisfy $\tau(y) < 1$. Then such a solution is unique.*

*Proof.* Let $y_1 \in \boldsymbol{V}$ and $y_2 \in \boldsymbol{V}$ be two solutions of (11.4). Set $Y = y_1 - y_2$. Then

$$\alpha\int_\Omega Y \cdot w + \nu\int_\Omega \nabla Y \cdot \nabla w + \int_\Omega y_2 \cdot \nabla Y \cdot w + \int_\Omega Y \cdot \nabla y_1 \cdot w = 0 \quad \forall w \in \boldsymbol{V}.$$

We now take $w = Y$ and use that $\int_\Omega y_2 \cdot \nabla Y \cdot Y = 0$. Using (11.12) and (11.13), we get

$$|||Y|||_{\boldsymbol{V}}^2 = -\int_\Omega Y \cdot \nabla y_1 \cdot Y \leq \tau(y_1)|||Y|||_{\boldsymbol{V}}^2,$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

leading to $(1-\tau(y_1))|||Y|||_{\boldsymbol{V}}^2 \le 0$. Consequently, if $\tau(y_1) < 1$, then $Y = 0$, and the solution of (11.4) is unique. In particular, in view of (11.14), this holds if the data satisfy $\nu\|g\|_2^2 + \frac{c_0}{\alpha}\|f\|_{H^{-1}(\Omega)^2}^2 < 2\nu^2$. □

We now introduce our least-squares functional $E : \boldsymbol{V} \to \mathbb{R}^+$ as

$$E(y) := \frac{1}{2}\int_\Omega \left(\alpha|v|^2 + \nu|\nabla v|^2\right) = \frac{1}{2}|||v|||_{\boldsymbol{V}}^2, \tag{11.15}$$

where the corrector $v \in \boldsymbol{V}$ is a unique solution of the linear formulation (11.9). In particular, the corrector $v$ satisfies the estimate

$$|||v|||_{\boldsymbol{V}} \le |||y|||_{\boldsymbol{V}}\left(1 + \frac{|||y|||_{\boldsymbol{V}}}{2\sqrt{\alpha}\nu}\right) + \sqrt{\frac{c_0\|f\|_{H^{-1}(\Omega)^2}^2}{\nu} + \alpha\|g\|_2^2}. \tag{11.16}$$

Conversely, we also have

$$|||y|||_{\boldsymbol{V}} \le |||v|||_{\boldsymbol{V}} + \sqrt{\frac{c_0\|f\|_{H^{-1}(\Omega)^2}^2}{\nu} + \alpha\|g\|_2^2}. \tag{11.17}$$

The infimum of $E$ is equal to zero and is reached by a solution of (11.4). In this sense the functional $E$ is a so-called error functional, which measures, through the corrector variable $v$, the deviation of the pair $y$ from being a solution of the underlying equation (11.4).

A practical way of taking a functional to its minimum is through some (clever) use of descent directions, i. e., the use of its derivative. In doing so, the presence of local minima is always something that may dramatically spoil the whole scheme. The unique structural property that discards this possibility is the strict convexity of the functional. However, for nonlinear equations like (11.4), we cannot expect this property to hold for the functional $E$ in (11.15). Nevertheless, we insist in that for a descent strategy applied to the extremal problem $\min_{y\in\boldsymbol{V}} E(y)$, numerical procedures cannot converge except to a global minimizer leading $E$ down to zero.

Indeed, we would like to show that the only critical points for $E$ correspond to solutions of (11.4). In such a case the search for a solution $y$ of (11.4) is reduced to the minimization of $E$.

For any $y \in \boldsymbol{V}$, we now look for a solution $Y_1 \in \boldsymbol{V}$ of the following equation:

$$\alpha\int_\Omega Y_1\cdot w + \nu\int_\Omega \nabla Y_1\cdot\nabla w + \int_\Omega (y\cdot\nabla Y_1 + Y_1\cdot\nabla y)\cdot w = -\alpha\int_\Omega v\cdot w - \nu\int_\Omega \nabla v\cdot\nabla w \quad \forall w \in \boldsymbol{V}, \tag{11.18}$$

where $v \in \boldsymbol{V}$ is the corrector (associated with $y$) solution of (11.9). The solution $Y_1$ enjoys the following property.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Proposition 11.2.5.** *For all $y \in \mathbb{B}$, there exists a unique solution $Y_1$ of* (11.18) *associated with $y$. Moreover, this solution satisfies*

$$(1-\tau(y))|||Y_1|||_{\boldsymbol{V}} \leq \sqrt{2E(y)}. \tag{11.19}$$

*Proof.* The proof uses the arguments of Lemma 11.2.4. We define the bilinear and continuous form $a : \boldsymbol{V} \times \boldsymbol{V} \to \mathbb{R}$ by

$$a(Y,w) = \alpha \int_\Omega Y \cdot w + \nu \int_\Omega \nabla Y \cdot \nabla w + \int_\Omega (y \cdot \nabla Y + Y \cdot \nabla y) \cdot w, \tag{11.20}$$

so that $a(Y,Y) = |||Y|||_{\boldsymbol{V}}^2 + \int_\Omega Y \cdot \nabla y \cdot Y$. Using (11.13), we obtain $a(Y,Y) \geq (1-\tau(y))|||Y|||_{\boldsymbol{V}}^2$ for all $Y \in \boldsymbol{V}$. The Lax–Milgram lemma leads to the existence and uniqueness of $Y_1$, provided that $\tau(y) < 1$. Then, putting $w = Y_1$ into (11.18) implies

$$a(Y_1,Y_1) \leq -\alpha \int_\Omega v \cdot Y_1 - \nu \int_\Omega \nabla v \cdot \nabla Y_1 \leq |||Y_1|||_{\boldsymbol{V}} |||v|||_{\boldsymbol{V}} = |||Y_1|||_{\boldsymbol{V}} \sqrt{2E(y)},$$

leading to (11.19). □

We now check the differentiability of the least-squares functional.

**Proposition 11.2.6.** *For all $y \in \boldsymbol{V}$, the map $Y \mapsto E(y+Y)$ is a differentiable function on the Hilbert space $\boldsymbol{V}$, and for any $Y \in \boldsymbol{V}$, we have*

$$E'(y) \cdot Y = \int_\Omega \alpha v \cdot V + \nu \nabla v \cdot \nabla V, \tag{11.21}$$

*where $V \in \boldsymbol{V}$ is the unique solution of*

$$\alpha \int_\Omega V \cdot w + \nu \int_\Omega \nabla V \cdot \nabla w = -\alpha \int_\Omega Y \cdot w - \nu \int_\Omega \nabla Y \cdot \nabla w - \int_\Omega (y \cdot \nabla Y + Y \cdot \nabla y) \cdot w \quad \forall w \in \boldsymbol{V}. \tag{11.22}$$

*Proof.* Let $y \in \boldsymbol{V}$ and $Y \in \boldsymbol{V}$. We have $E(y+Y) = \frac{1}{2}|||\overline{V}|||_{\boldsymbol{V}}^2$, where $\overline{V} \in \boldsymbol{V}$ is the unique solution of

$$\begin{aligned} &\alpha \int_\Omega \overline{V} \cdot w + \nu \int_\Omega \nabla \overline{V} \cdot \nabla w + \alpha \int_\Omega (y+Y) \cdot w + \nu \int_\Omega \nabla (y+Y) \cdot \nabla w \\ &\quad + \int_\Omega (y+Y) \cdot \nabla (y+Y) \cdot w - \langle f, w \rangle_{\boldsymbol{V}' \times \boldsymbol{V}} - \alpha \int_\Omega g w = 0 \quad \forall w \in \boldsymbol{V}. \end{aligned}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

If $v \in \boldsymbol{V}$ is the solution of (11.9) associated with $y$, then $v' \in \boldsymbol{V}$ is the unique solution of

$$\alpha \int_\Omega v' \cdot w + \nu \int_\Omega \nabla v' \cdot \nabla w + \int_\Omega Y \cdot \nabla Y \cdot w = 0 \quad \forall w \in \boldsymbol{V}, \tag{11.23}$$

and $V \in \boldsymbol{V}$ is the unique solution of (11.22), then it is straightforward to check that $\overline{V} - v - v' - V \in \boldsymbol{V}$ is solution of

$$\alpha \int_\Omega (\overline{V} - v - v' - V) \cdot w + \nu \int_\Omega \nabla(\overline{V} - v - v' - V) \cdot \nabla w = 0 \quad \forall w \in \boldsymbol{V},$$

and therefore $\overline{V} - v - v' - V = 0$. Thus

$$\begin{aligned} E(y+Y) =& \frac{1}{2}|||v + v' + V|||_{\boldsymbol{V}}^2 \\ =& \frac{1}{2}|||v|||_{\boldsymbol{V}}^2 + \frac{1}{2}|||v'|||_{\boldsymbol{V}}^2 + \frac{1}{2}|||V|||_{\boldsymbol{V}}^2 + \langle V, v' \rangle_{\boldsymbol{V}} + \langle V, v \rangle_{\boldsymbol{V}} + \langle v, v' \rangle_{\boldsymbol{V}}. \end{aligned} \tag{11.24}$$

Then, writing (11.22) with $w = V$ and using (11.12), we obtain

$$\begin{aligned} |||V|||_{\boldsymbol{V}}^2 &\le |||V|||_{\boldsymbol{V}} |||Y|||_{\boldsymbol{V}} + \sqrt{2}(\|y\|_2 \|\nabla Y\|_2 + \|Y\|_2 \|\nabla y\|_2)\|\nabla V\|_2 \\ &\le |||V|||_{\boldsymbol{V}} |||Y|||_{\boldsymbol{V}} + \frac{\sqrt{2}}{\sqrt{\alpha\nu}} |||y|||_{\boldsymbol{V}} |||Y|||_{\boldsymbol{V}} \|\nabla V\|_2, \end{aligned}$$

leading to $|||V|||_{\boldsymbol{V}} \le |||Y|||_{\boldsymbol{V}}(1 + \frac{\sqrt{2}}{\sqrt{\alpha\nu}}|||y|||_{\boldsymbol{V}})$. Similarly, using (11.23), we obtain $|||v'|||_{\boldsymbol{V}} \le \frac{1}{\sqrt{2\alpha\nu}}|||Y|||_{\boldsymbol{V}}^2$. It follows that $\frac{1}{2}|||v'|||_{\boldsymbol{V}}^2 + \frac{1}{2}|||V|||_{\boldsymbol{V}}^2 + \langle V, v' \rangle_{\boldsymbol{V}} + \langle v, v' \rangle_{\boldsymbol{V}} = o(|||Y|||_{\boldsymbol{V}})$ and from (11.24) that

$$E(y+Y) = E(y) + \langle v, V \rangle + o(|||Y|||_{\boldsymbol{V}}).$$

Eventually, the estimate $|\langle v, V \rangle_{\boldsymbol{V}}| \le |||v|||_{\boldsymbol{V}} |||V|||_{\boldsymbol{V}} \le (1 + \frac{\sqrt{2}}{\sqrt{\alpha\nu}}|||y|||_{\boldsymbol{V}})\sqrt{E(y)}|||Y|||_{\boldsymbol{V}}$ gives the continuity of the linear map $Y \mapsto \langle v, V \rangle_{\boldsymbol{V}}$. □

We are now in position to prove the following result, which indicates that in the ball $\mathbb{B}$ of $\boldsymbol{V}$, any critical point for $E$ is also a zero of $E$.

**Proposition 11.2.7.** *For all $y \in \mathbb{B}$,*

$$(1 - \tau(y))\sqrt{2E(y)} \le \frac{1}{\sqrt{\nu}} \|E'(y)\|_{\boldsymbol{V}'}.$$

*Proof.* For any $Y \in \boldsymbol{V}$, $E'(y) \cdot Y = \int_\Omega \alpha v \cdot V + \nu \nabla v \cdot \nabla V$, where $V \in \boldsymbol{V}$ is the unique solution of (11.22). In particular, taking $Y = Y_1$ defined by (11.18), we obtain a solution

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$V_1 \in \boldsymbol{V}$ of

$$\alpha \int_\Omega V_1 \cdot w + \nu \int_\Omega \nabla V_1 \cdot \nabla w = -\alpha \int_\Omega Y_1 \cdot w - \nu \int_\Omega \nabla Y_1 \cdot \nabla w - \int_\Omega (y \cdot \nabla Y_1 + Y_1 \cdot \nabla y) \cdot w \quad \forall w \in \boldsymbol{V}. \tag{11.25}$$

Summing (11.18) and (11.25), we obtain that $v - V_1 \in \boldsymbol{V}$ solves

$$\alpha \int_\Omega (v - V_1) w + \nu \int_\Omega (\nabla v - \nabla V_1) \cdot w = 0 \quad \forall w \in \boldsymbol{V}.$$

This implies that $v$ and $V_1$ coincide and then that

$$E'(y) \cdot Y_1 = \int_\Omega \alpha |v|^2 + \nu |\nabla v|^2 = 2E(y) \quad \forall y \in \boldsymbol{V}. \tag{11.26}$$

It follows that $2E(y) = E'(y) \cdot Y_1 \leq \|E'(y)\|_{\boldsymbol{V}'} \|Y_1\|_{\boldsymbol{V}} \leq \|E'(y)\|_{\boldsymbol{V}'} \frac{\|Y_1\|_V}{\sqrt{\nu}}$. Proposition 11.2.5 allows us to conclude. □

Eventually, we prove the following coercivity-type inequality for the error functional $E$.

**Proposition 11.2.8.** *Assume that a solution $\overline{y} \in \boldsymbol{V}$ of* (11.4) *satisfies $\tau(\overline{y}) < 1$. Then, for all $y \in \boldsymbol{V}$,*

$$|||y - \overline{y}|||_{\boldsymbol{V}} \leq \left(1 - \tau(\overline{y})\right)^{-1} \sqrt{2E(y)}. \tag{11.27}$$

*Proof.* For any $y \in \boldsymbol{V}$, let $v$ be the corresponding corrector, and let $Y = y - \overline{y}$. We have

$$\alpha \int_\Omega Y \cdot w + \nu \int_\Omega \nabla Y \cdot \nabla w + \int_\Omega y \cdot \nabla Y \cdot w + \int_\Omega Y \cdot \nabla \overline{y} \cdot w = -\alpha \int_\Omega v \cdot w - \nu \int_\Omega \nabla v \cdot \nabla w \quad \forall w \in \boldsymbol{V}. \tag{11.28}$$

For $w = Y$, this equality rewrites

$$|||Y|||_{\boldsymbol{V}}^2 = -\int_\Omega Y \cdot \nabla \overline{y} \cdot Y - \alpha \int_\Omega v \cdot Y - \nu \int_\Omega \nabla v \cdot \nabla Y.$$

The results follow by repeating the arguments of the proof of Proposition 11.2.5. □

Assuming the existence of a solution of (11.4) in the ball $\mathbb{B}$, Propositions 11.2.7 and 11.2.8 imply that any minimizing sequence $(y_k)_{(k \in \mathbb{N})}$ for $E$ uniformly in $\mathbb{B}$ strongly converges to a solution of (11.4). Remark that by Lemma 11.2.4 such a solution is unique. In the next section, we construct such a sequence $(y_k)_{(k \in \mathbb{N})}$, assuming the parameter $\alpha$ to be large enough.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Remark 11.2.9.** To simplify the notations, we have introduced the corrector variable $v$ leading to the functional $E$. Instead, we may consider the functional $\widetilde{E} : \boldsymbol{V} \to \mathbb{R}$ defined by

$$\widetilde{E}(y) := \frac{1}{2}\|\alpha y + \nu B_1(y) + B(y,y) - f + \alpha g\|^2_{\boldsymbol{V}'}$$

with $B_1 : \boldsymbol{V} \to L^2(\Omega)^2$ and $B : \boldsymbol{V} \times \boldsymbol{V} \to L^2(\Omega)^2$ defined by $(B_1(y), w) := (\nabla y, \nabla w)_2$ and $(B(y,z), w) := \int_\Omega y \cdot \nabla z \cdot w$, respectively. The functionals $E$ and $\widetilde{E}$ are equivalent. Precisely, from the definition of $v$ (see (11.9)) we deduce that

$$E(y) = \frac{1}{2}|||v|||^2_{\boldsymbol{V}} \le \frac{c_0^2}{2\nu^2}\|\alpha y + \nu B_1(y) + B(y,y) - f + \alpha g\|^2_{\boldsymbol{V}'} = \frac{c_0^2}{\nu^2}\widetilde{E}(y) \quad \forall y \in \boldsymbol{V}.$$

Conversely,

$$\begin{aligned}\|\alpha y + \nu B_1(y) + B(y,y) - f + \alpha g\|_{\boldsymbol{V}'} &= \sup_{w\in\boldsymbol{V}, w\neq 0} \frac{\int_\Omega (\alpha v w + \nu \nabla v \cdot \nabla w)}{\|w\|_{\boldsymbol{V}}} \\ &\le |||v|||_{\boldsymbol{V}} \sup_{w\in\boldsymbol{V}, w\neq 0} \frac{|||w|||_{\boldsymbol{V}}}{\|w\|_{\boldsymbol{V}}} \le \sqrt{\alpha c_0^2 + \nu}\,|||v|||_{\boldsymbol{V}},\end{aligned}$$

so that $\widetilde{E}(y) \le (\alpha c_0^2 + \nu)E(y)$ for all $y \in \boldsymbol{V}$.

### 11.2.2 A strongly convergent minimizing sequence for $E$-link with the damped Newton method

Iin this section, we define a sequence converging strongly to a solution of (11.4) for which $E$ vanishes. According to Proposition 11.2.7, it suffices to define a minimizing sequence for $E$ included in the ball $\mathbb{B} = \{y \in \boldsymbol{V}, \tau(y) < 1\}$. In this respect, remark that equality (11.26) shows that $-Y_1$ given by the solution of (11.18) is a descent direction for the functional $E$. Therefore for any $m \ge 1$, we can define, at least formally, a minimizing sequence $\{y_k\}_{(k\ge 0)}$ as follows:

$$\begin{cases} y_0 \in \boldsymbol{H} \quad \text{given}, \\ y_{k+1} = y_k - \lambda_k Y_{1,k}, \quad k \ge 0, \\ \lambda_k = \underset{\lambda\in[0,m]}{\arg\min}\, E(y_k - \lambda Y_{1,k}) \end{cases} \tag{11.29}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

with the solution $Y_{1,k} \in \boldsymbol{V}$ of the equation

$$\begin{aligned} \alpha \int_\Omega Y_{1,k} \cdot w + \nu \int_\Omega \nabla Y_{1,k} \cdot \nabla w + \int_\Omega (y_k \cdot \nabla Y_{1,k} + Y_{1,k} \cdot \nabla y_k) \cdot w \\ = -\alpha \int_\Omega v_k \cdot w - \nu \int_\Omega \nabla v_k \cdot \nabla w \quad \forall w \in \boldsymbol{V} \end{aligned} \tag{11.30}$$

and the corrector solution $v_k \in \boldsymbol{V}$ (associated with $y_k$) of (11.9), leading (see (11.26)) to $E'(y_k) \cdot Y_{1,k} = 2E(y_k)$.

Remark that by (11.17) the sequence $(y_k)_{k>0}$ is uniformly bounded since $y_k$ satisfies $|||y_k|||_{\boldsymbol{V}} \leq \sqrt{2E(y_k)} + \sqrt{\frac{c_0}{\nu}\|f\|^2_{H^{-1}} + \alpha\|g\|_2^2}$. However, we insist that to justify the existence of the element $Y_{1,k}$, $y_k$ should satisfy $\tau(y_k) < 1$, i. e., $\|\nabla y_k\|_2 < \sqrt{2\alpha\nu}$. We proceed in two steps. First, assuming that the sequence $\{y_k\}_{(k>0)}$ defined by (11.29) satisfies $\tau(y_k) \leq c_1 < 1$ for any $k$, we show that $E(y_k) \to 0$ and that $(y_k)_k$ converges strongly in $\boldsymbol{V}$ to a solution of (11.4). Then we determine sufficient conditions on the initial guess $y_0 \in \boldsymbol{V}$ in order that $\tau(y_k) < 1$ for all $k \in \mathbb{N}$.

We start with the following lemma, which provides the main property of the sequence $(E(y_k))_{(k\geq 0)}$.

**Lemma 11.2.10.** *Assume that the sequence $(y_k)_{(k\geq 0)}$ defined by* (11.29) *satisfy $\tau(y_k) < 1$. Then, for all $\lambda \in \mathbb{R}$, we have the following estimate:*

$$E(y_k - \lambda Y_{1,k}) \leq E(y_k)\left(|1-\lambda| + \lambda^2 \frac{(1-\tau(y_k))^{-2}}{\sqrt{\alpha\nu}}\sqrt{E(y_k)}\right)^2. \tag{11.31}$$

*Proof.* For any real $\lambda$ and any $y_k, w_k \in \boldsymbol{V}$, we get the following expansion:

$$\begin{aligned} E(y_k - \lambda w_k) = E(y_k) - \lambda \int_\Omega (\alpha v_k \overline{v}_k + \nu \nabla v_k \cdot \nabla \overline{v}_k) \\ + \frac{\lambda^2}{2}\int_\Omega \left(\alpha|\overline{v}_k|^2 + \nu|\nabla \overline{v}_k|^2 + 2(\alpha v_k \overline{\overline{v}}_k + \nu \nabla v_k \cdot \nabla \overline{\overline{v}}_k)\right) \\ - \lambda^3 \int_\Omega \alpha \overline{v}_k \overline{\overline{v}}_k + \nu \nabla \overline{v}_k \cdot \nabla \overline{\overline{v}}_k + \frac{\lambda^4}{2}\int_\Omega \alpha|\overline{\overline{v}}_k|^2 + \nu|\nabla \overline{\overline{v}}_k|^2, \end{aligned} \tag{11.32}$$

where $v_k, \overline{v}_k \in \boldsymbol{V}$ and $\overline{\overline{v}}_k \in \boldsymbol{V}$ solve, respectively,

$$\begin{aligned} \alpha \int_\Omega v_k \cdot w + \nu \int_\Omega \nabla v_k \cdot \nabla w + \alpha \int_\Omega y_k \cdot w + \nu \int_\Omega \nabla y_k \cdot \nabla w + \int_\Omega y_k \cdot \nabla y_k \cdot w \\ = \langle f, w\rangle_{H^{-1}(\Omega)^2 \times H^1_0(\Omega)^2} + \alpha \int_\Omega g \cdot w \quad \forall w \in \boldsymbol{V}, \end{aligned} \tag{11.33}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$$\begin{aligned}\alpha \int_\Omega \overline{v}_k \cdot w + \nu \int_\Omega \nabla \overline{v}_k \cdot \nabla w + \alpha \int_\Omega w_k \cdot w + \nu \int_\Omega \nabla w_k \cdot \nabla w \\ + \int_\Omega w_k \cdot \nabla y_k \cdot w + y_k \cdot \nabla w_k \cdot w = 0 \quad \forall w \in \boldsymbol{V},\end{aligned} \tag{11.34}$$

and

$$\alpha \int_\Omega \overline{\overline{v}}_k \cdot w + \nu \int_\Omega \nabla \overline{\overline{v}}_k \cdot \nabla w + \int_\Omega w_k \cdot \nabla w_k \cdot w = 0 \quad \forall w \in \boldsymbol{V}. \tag{11.35}$$

Since the corrector $\overline{v}_k$ associated with $Y_{1,k}$ coincides with the corrector $v_k$ associated with $y_k$, expansion (11.32) reduces to

$$\begin{aligned}E(y_k - \lambda Y_{1,k}) &= (1-\lambda)^2 E(y_k) + \lambda^2(1-\lambda) \int_\Omega \alpha v_k \overline{\overline{v}}_k + \nu \nabla v_k \nabla \overline{\overline{v}}_k \\ &\quad + \frac{\lambda^4}{2} \int_\Omega \alpha |\overline{\overline{v}}_k|^2 + \nu |\nabla \overline{\overline{v}}_k|^2 \\ &\leq (1-\lambda)^2 E(y_k) + \lambda^2(1-\lambda) |||v_k|||_{\boldsymbol{V}} |||\overline{\overline{v}}_k|||_{\boldsymbol{V}} + \frac{\lambda^4}{2} |||\overline{\overline{v}}_k|||_{\boldsymbol{V}}^2 \\ &\leq \left( |1-\lambda| \sqrt{E(y_k)} + \frac{\lambda^2}{\sqrt{2}} |||\overline{\overline{v}}_k|||_{\boldsymbol{V}} \right)^2.\end{aligned} \tag{11.36}$$

Equality (11.35) then leads to $|||\overline{\overline{v}}_k|||_{\boldsymbol{V}} \leq \frac{|||Y_{1,k}|||_{\boldsymbol{V}}^2}{\sqrt{2\alpha\nu}} \leq \sqrt{2}(1-\tau(y_k))^{-2} \frac{E(y_k)}{\sqrt{\alpha\nu}}$ and to (11.31). □

We are now in position to prove the following convergence result for the sequence $\{E(y_k)\}_{(k\geq 0)}$.

**Proposition 11.2.11.** *Let $(y_k)_{k\geq 0}$ be the sequence defined by* (11.29). *Assume that there exists a constant $c_1 \in (0,1)$ such that $\tau(y_k) \leq c_1$ for all $k$. Then $E(y_k) \to 0$ as $k \to \infty$. Moreover, there exists $k_0 \in \mathbb{N}$ such that the sequence $(E(y_k))_{(k\geq k_0)}$ decays quadratically.*

*Proof.* The inequality $\tau(y_k) \leq c_1$ and (11.31) imply that

$$E(y_k - \lambda Y_{1,k}) \leq E(y_k)\left(|1-\lambda| + \lambda^2 c_{\alpha,\nu} \sqrt{E(y_k)}\right)^2, \quad c_{\alpha,\nu} := \frac{(1-c_1)^{-2}}{\sqrt{\alpha\nu}}. \tag{11.37}$$

Let us define the real function $p_k(\lambda) = |1-\lambda| + \lambda^2 c_{\alpha,\nu} \sqrt{E(y_k)}$ for $\lambda \in [0,m]$. So we can write

$$\sqrt{E(y_{k+1})} = \min_{\lambda\in[0,m]} \sqrt{E(y_k - \lambda Y_{1,k})} \leq \min_{\lambda\in[0,m]} p_k(\lambda) \sqrt{E(y_k)}.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

If $c_{\alpha,\nu}\sqrt{E(y_0)} < 1$ (and thus $c_{\alpha,\nu}\sqrt{E(y_k)} < 1$ for all $k \in \mathbb{N}$), then

$$p_k(\widetilde{\lambda}_k) := \min_{\lambda\in[0,m]} p_k(\lambda) \le p_k(1) = c_{\alpha,\nu}\sqrt{E(y_k)},$$

and thus

$$c_{\alpha,\nu}\sqrt{E(y_{k+1})} \le \big(c_{\alpha,\nu}\sqrt{E(y_k)}\big)^2, \tag{11.38}$$

implying that $c_{\alpha,\nu}\sqrt{E(y_k)} \to 0$ as $k \to \infty$ with quadratic rate.

Suppose now that $c_{\alpha,\nu}\sqrt{E(y_0)} \ge 1$ and denote $I = \{k \in \mathbb{N},\ c_{\alpha,\nu}\sqrt{E(y_k)} \ge 1\}$. Let us prove that $I$ is a finite subset of $\mathbb{N}$. For all $k \in I$, since $c_{\alpha,\nu}\sqrt{E(y_k)} \ge 1$,

$$\min_{\lambda\in[0,m]} p_k(\lambda) = \min_{\lambda\in[0,1]} p_k(\lambda) = p_k\bigg(\frac{1}{2c_{\alpha,\nu}\sqrt{E(y_k)}}\bigg) = 1 - \frac{1}{4c_{\alpha,\nu}\sqrt{E(y_k)}},$$

and thus, for all $k \in I$,

$$c_{\alpha,\nu}\sqrt{E(y_{k+1})} \le \bigg(1 - \frac{1}{4c_{\alpha,\nu}\sqrt{E(y_k)}}\bigg)c_{\alpha,\nu}\sqrt{E(y_k)} = c_{\alpha,\nu}\sqrt{E(y_k)} - \frac{1}{4}.$$

This inequality implies that the sequence $(c_{\alpha,\nu}\sqrt{E(y_k)})_{k\in\mathbb{N}}$ strictly decreases, and thus there exists $k_0 \in \mathbb{N}$ such that $c_{\alpha,\nu}\sqrt{E(y_k)} < 1$ for all $k \ge k_0$. Thus $I$ is a finite subset of $\mathbb{N}$. Arguing as in the first case, it follows that $c_{\alpha,\nu}\sqrt{E(y_k)} \to 0$ as $k \to \infty$.

In both cases, remark that $p_k(\widetilde{\lambda}_k)$ decreases with respect to $k$. □

**Lemma 11.2.12.** *Assume that the sequence $(y_k)_{(k\ge0)}$ defined by* (11.29) *satisfies $\tau(y_k) \le c_1$ for all $k$ and some $c_1 \in (0,1)$. Then $\lambda_k \to 1$ as $k \to \infty$.*

*Proof.* In view of (11.36), we have, as long as $E(y_k) > 0$,

$$(1-\lambda_k)^2 = \frac{E(y_{k+1})}{E(y_k)} - \lambda_k^2(1-\lambda_k)\frac{\langle v_k, \overline{\overline{v}}_k\rangle_{\boldsymbol{V}}}{E(y_k)} - \lambda_k^4\frac{|||\overline{\overline{v}}_k|||_{\boldsymbol{V}}^2}{2E(y_k)}.$$

From the proof of Lemma 11.2.10, $\frac{\langle v_k,\overline{\overline{v}}_k\rangle_{\boldsymbol{V}}}{E(y_k)} \le C(\alpha,\nu)(1-c_1)^{-2}\sqrt{E(y_k)}$, whereas $\frac{|||\overline{\overline{v}}_k|||_{\boldsymbol{V}}^2}{E(y_k)} \le C(\alpha,\nu)^2(1-c_1)^{-4}E(y_k)$. Consequently, since $\lambda_k \in [0,m]$ and $\frac{E(y_{k+1})}{E(y_k)} \to 0$, we deduce that $(1-\lambda_k)^2 \to 0$, that is, $\lambda_k \to 1$ as $k \to \infty$. □

**Proposition 11.2.13.** *Let $(y_k)_{k\in\mathbb{N}}$ be the sequence defined by* (11.29). *Assume that there exists a constant $c_1 \in (0,1)$ such that $\tau(y_k) \le c_1$ for all $k$. Then $y_k \to \overline{y}$ in $\boldsymbol{V}$, where $\overline{y} \in \boldsymbol{V}$ is the unique solution of* (11.4).

*Proof.* Remark that we cannot use Proposition 11.2.8 since we do not know yet that there exists a solution, say $z$, of (11.4) satisfying $\tau(z) < 1$. In view of $y_{k+1} = y_0 -$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$\sum_{n=0}^{k} \lambda_n Y_{1,n}$, we write

$$\sum_{n=0}^{k} |\lambda_n| |||Y_{1,n}|||_{\boldsymbol{V}} \le m \sum_{n=0}^{k} |||Y_{1,n}|||_{\boldsymbol{V}} \le m\sqrt{2} \sum_{n=0}^{k} \frac{\sqrt{E(y_n)}}{1-\tau_d(y_n)} \le \frac{m\sqrt{2}}{1-c_1} \sum_{n=0}^{k} \sqrt{E(y_n)}.$$

Using that $p_n(\widetilde{\lambda}_n) \le p_0(\widetilde{\lambda}_0)$ for all $n \ge 0$, we can write, for $n > 0$,

$$\sqrt{E(y_n)} \le p_{n-1}(\widetilde{\lambda}_{n-1}) \sqrt{E(y_{n-1})} \le p_0(\widetilde{\lambda}_0) \sqrt{E(y_{n-1})} \le p_0(\widetilde{\lambda}_0)^n \sqrt{E(y_0)}.$$

Recalling that $p_0(\widetilde{\lambda}_0) = \min_{\lambda\in[0,1]} p_0(\lambda) < 1$ since $p_0(0) = 1$ and $p_0'(0) = -1$, we finally obtain

$$\sum_{n=0}^{k} |\lambda_n| |||Y_{1,n}|||_{\boldsymbol{V}} \le \frac{m\sqrt{2}}{1-c_1} \frac{\sqrt{E(y_0)}}{1-p_0(\widetilde{\lambda}_0)},$$

for which we deduce that the series $\sum_{k\ge0} \lambda_k Y_{1,k}$ converges in $\boldsymbol{V}$. Then $y_k$ converges in $\boldsymbol{V}$ to $\overline{y} := y_0 + \sum_{k\ge0} \lambda_k Y_{1,k}$. Eventually, the convergence of $E(y_k)$ to 0 implies the convergence of the corrector $v_k$ to 0 in $\boldsymbol{V}$; taking the limit in the corrector equation (11.33) shows that $\overline{y}$ solves (11.4). Since $\tau(\overline{y}) \le c_1 < 1$, Lemma 11.2.4 shows that this solution is unique. □

As mentioned earlier, the remaining and crucial point is to show that the sequence $\{y_k\}$ may satisfy the uniform property $\tau(y_k) \le c_1$ for some $c_1 < 1$.

**Lemma 11.2.14.** *Let $y_0 = g \in \boldsymbol{V}$. For all $c_1 \in (0,1)$, there exists $\alpha_0 > 0$ such that, for any $\alpha \ge \alpha_0$, the unique sequence defined by* (11.29) *satisfies $\tau(y_k) \le c_1$ for all $k \ge 0$.*

*Proof.* Let $c_1 \in (0,1)$ and assume that $y_0$ belongs to $\boldsymbol{V}$. There exists $\alpha_1 > 0$ such that $\tau(y_0) \le \frac{c_1}{2}$ for all $\alpha \ge \alpha_1$.

Moreover, in view of the above computation, since $\|v\|_{\boldsymbol{V}} \le \frac{1}{\nu} |||v|||_{\boldsymbol{V}}$ for all $v \in \boldsymbol{V}$, we have, for all $\alpha > 0$ and $k \in \mathbb{N}$,

$$\|y_{k+1}\|_{\boldsymbol{V}} \le \|y_0\|_{\boldsymbol{V}} + \frac{m\sqrt{2}}{\nu(1-c_1)} \frac{\sqrt{E(y_0)}}{1-p_0(\widetilde{\lambda}_0)},$$

where

$$\frac{\sqrt{E(y_0)}}{1-p_0(\widetilde{\lambda}_0)} \le \begin{cases} \frac{\sqrt{E(y_0)}}{1-c_{\alpha,\nu}\sqrt{E(y_0)}} & \text{if } c_{\alpha,\nu}\sqrt{E(y_0)} < 1, \\ 4c_{\alpha,\nu}E(y_0) & \text{if } c_{\alpha,\nu}\sqrt{E(y_0)} \ge 1. \end{cases}$$

From (11.9) we obtain that

$$|||v|||_{\boldsymbol{V}}^2 \le \alpha \|g-y\|_2^2 + \frac{1}{\nu}\big(\nu\|\nabla y\|_2 + \|y\|_2\|\nabla y\|_2 + \sqrt{c_0}\|f\|_{H^{-1}(\Omega)^2}\big)^2.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

In particular, taking $y = y_0 = g$ allows us to remove the $\alpha$ term in the right-hand side and gives

$$E(g) \leq \frac{1}{2\nu}\left(\|g\|_{\boldsymbol{V}}(\nu + \|g\|_2) + \sqrt{c_0}\|f\|_{H^{-1}(\Omega)^2}\right)^2 := \frac{1}{2\nu}c_2(f,g), \tag{11.39}$$

and thus, if $c_{\alpha_1,\nu}\sqrt{E(g)} \geq 1$, then for all $\alpha \geq \alpha_1$ such that $c_{\alpha,\nu}\sqrt{E(g)} \geq 1$ and for all $k \in \mathbb{N}$,

$$\|y_{k+1}\|_{\boldsymbol{V}} \leq \|g\|_{\boldsymbol{V}} + \frac{m\sqrt{2}}{\nu(1-c_1)}\frac{\sqrt{E(g)}}{1-p_0(\tilde{\lambda}_0)} \leq \|g\|_{\boldsymbol{V}} + \frac{2m\sqrt{2}}{\nu^3\sqrt{\alpha}(1-c_1)^3}c_2(f,g). \tag{11.40}$$

If now $c_{\alpha_1,\nu}\sqrt{E(g)} < 1$, then there exists $0 < K < 1$ such that for all $\alpha \geq \alpha_1$, we have $c_{\alpha,\nu}\sqrt{E(g)} \leq K$. Therefore, for all $\alpha \geq \alpha_1$, we have

$$\frac{\sqrt{E(g)}}{1-p_0(\tilde{\lambda}_0)} \leq \frac{\sqrt{E(g)}}{1-K},$$

and thus, for all $k \in \mathbb{N}$,

$$\|y_{k+1}\|_{\boldsymbol{V}} \leq \|g\|_{\boldsymbol{V}} + \frac{m\sqrt{2}}{\nu(1-c_1)}\frac{\sqrt{E(g)}}{1-p_0(\tilde{\lambda}_0)} \leq \|g\|_{\boldsymbol{V}} + \frac{m}{\nu^{3/2}(1-c_1)(1-K)}\sqrt{c_2(f,g)}. \tag{11.41}$$

On the other hand, there exists $\alpha_0 \geq \alpha_1$ such that, for all $\alpha \geq \alpha_0$, we have

$$\frac{2m\sqrt{2}}{\nu^3\sqrt{\alpha}(1-c_1)^3}c_2(f,g) \leq \frac{c_1}{2}\sqrt{2\alpha\nu}$$

and

$$\frac{m}{\nu^{3/2}(1-c_1)(1-K)}\sqrt{c_2(f,g)} \leq \frac{c_1}{2}\sqrt{2\alpha\nu}.$$

We then deduce from (11.40) and (11.41) that for all $\alpha \geq \alpha_0$ and all $k \in \mathbb{N}$,

$$\|y_{k+1}\|_{\boldsymbol{V}} \leq \frac{c_1}{2}\sqrt{2\alpha\nu} + \frac{c_1}{2}\sqrt{2\alpha\nu} = c_1\sqrt{2\alpha\nu},$$

that is, $\tau(y_{k+1}) \leq c_1$. □

Gathering the previous lemmas and propositions, we can now deduce the strong convergence of the sequence $(y_k)_{k\geq 0}$ defined by (11.29) and initialized by $y_0 = g$.

**Theorem 11.2.15.** *Let $c_1 \in (0,1)$. Assume that $y_0 = g \in \boldsymbol{V}$ and $\alpha$ is large enough so that*

$$c_2(f,g) \leq \max\left(\frac{1-c_1}{2}, \frac{c_1\sqrt{\nu}(1-K^2)}{m}\right)\frac{c_1}{4m}\nu^{5/2}(1-c_1)^2 2\alpha\nu. \tag{11.42}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*Then the sequence* $(y_k)_{(k\in\mathbb{N})}$ *defined by* (11.29) *strongly converges to the unique solution* $y$ *of* (11.4). *Moreover, there exists* $k_0 \in \mathbb{N}$ *such that the sequence* $(y_k)_{k\geq k_0}$ *converges quadratically to* $y$. *Moreover, this solution satisfies* $\tau(y) < 1$.

### 11.2.3 Remarks

**Remark 11.2.16.** Estimate (11.14) is usually used to obtain a sufficient condition on the data $f$, $g$ to ensure the uniqueness of the solution of (11.4) (i. e., $\tau(y) < 1$): this leads to

$$\alpha\|g\|_2^2 + \frac{c_0}{\nu}\|f\|^2_{(H^{-1}(\Omega))^2} \leq 2\alpha\nu. \tag{11.43}$$

We emphasize that such (sufficient) conditions are more restrictive than (11.42), as they impose smallness properties on $g$: precisely, $\|g\|_2^2 \leq 2\nu$. In particular, the latter yields a restrictive condition for large $\alpha$, contrary to (11.42).

**Remark 11.2.17.** It seems surprising that algorithm (11.29) achieves a quadratic rate for large $k$. Let us consider the application $\mathcal{F} : \boldsymbol{V} \to \boldsymbol{V}'$ defined as $\mathcal{F}(y) = \alpha y + \nu B_1(y) + B(y,y) - f - \alpha g$. The sequence $\{y_k\}_{k>0}$ associated to the Newton method to find the zero of $F$ is defined as follows:

$$\begin{cases} y_0 \in \boldsymbol{V}, \\ \mathcal{F}'(y_k)\cdot(y_{k+1} - y_k) = -\mathcal{F}(y_k), \quad k \geq 0. \end{cases} \tag{11.44}$$

We check that this sequence coincides with the sequence obtained from (11.29) if $\lambda_k$ is fixed equal to one and if $y_0 \in \boldsymbol{V}$. Algorithm (11.29), which optimizes the parameter $\lambda_k \in [0, m]$, $m \geq 1$, to minimize $E(y_k)$ or, equivalently, $\|\mathcal{F}(y_k)\|_{\boldsymbol{V}'}$, corresponds to the so-called in the literature damped Newton method for the application $\mathcal{F}$ (see [4]). As the iterates increase, the optimal parameter $\lambda_k$ converges to one (according to Lemma 11.2.12), and this globally convergent method behaves like the standard Newton method (for which $\lambda_k$ is fixed equal to one). This explains the quadratic rate after a finite number of iterates. To the best of our knowledge, this is the first analysis of the damped Newton method for partial differential equations. Among the few numerical works devoted to the damped Newton method for partial differential equations, we mention [21] for computing viscoplastic fluid flows.

**Remark 11.2.18.** Section 6 of Chapter 6 of the book [6] introduces a least-squares method to solve an Oseen-type equation (without incompressibility constraint). The convergence of any minimizing sequence toward a solution $y$ is proved under the a priori assumption that the operator $DF(y)$ defined as

$$DF(y)\cdot w = \alpha w - \nu\Delta w + \left[(w\cdot\nabla)y + (y\cdot\nabla)w\right] \quad \forall w \in \boldsymbol{V} \tag{11.45}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

(for some $\alpha > 0$) is an isomorphism from $\boldsymbol{V}$ onto $\boldsymbol{V}'$. Then $y$ is said to be a nonsingular point. According to Proposition 11.2.5, a sufficient condition for $y$ to be a nonsingular point is $\tau(y) < 1$. Recall that $\tau$ depends on $\alpha$. As far as we know, determining a weaker condition ensuring that $DF(y)$ is an isomorphism is an open question. Moreover, according to Lemma 11.2.4, it turns out that this condition is also a sufficient condition for the uniqueness of (11.4). Theorem 11.2.15 asserts that if $\alpha$ is large enough, then the sequence $(y_k)_{(k\in\mathbb{N})}$ defined in (11.29), initialized with $y_0 = g$, is a convergent sequence of nonsingular points. Since $\lambda_k$ is a constant equal to one, this shows the convergence of the Newton method to solve the steady Navier–Stokes equation.

**Remark 11.2.19.** We may also define a minimizing sequence for $E$ using the gradient $E'$:

$$\begin{cases} y_0 \in \boldsymbol{H} \quad \text{given}, \\ y_{k+1} = y_k - \lambda_k g_k, \quad k \geq 0, \\ \lambda_k = \underset{\lambda\in[0,m]}{\operatorname{argmin}} E(y_k - \lambda g_k) \end{cases} \tag{11.46}$$

with $g_k \in \boldsymbol{V}$ such that $(g_k, w)_{\boldsymbol{V}} = (E'(y_k), w)_{\boldsymbol{V}',\boldsymbol{V}}$ for all $w \in \boldsymbol{V}$. In particular, $\|g_k\|_{\boldsymbol{V}} = \|E'(y_k)\|_{\boldsymbol{V}'}$. Using expansion (11.24) with $w_k = g_k$, we can prove the linear decrease of the sequence $(E(y_k))_{k>0}$ to zero assuming, however, that $E(y_0)$ is small enough, of the order of $\nu^2$, independently of the value of $\alpha$.

## 11.2.4 Application to the backward Euler scheme

We use the analysis of the previous section to discuss the resolution of the backward Euler scheme (11.3) through a least-squares method. The weak formulation of this scheme reads as follows: given $y^0 = u_0 \in \boldsymbol{H}$, the sequence $\{y^n\}_{n>0}$ in $\boldsymbol{V}$ is defined by recurrence as follows:

$$\int_\Omega \frac{y^{n+1} - y^n}{\delta t} \cdot w + \nu \int_\Omega \nabla y^{n+1} \cdot \nabla w + \int_\Omega y^{n+1} \cdot \nabla y^{n+1} \cdot w = \langle f^n, w \rangle_{H^{-1}(\Omega)^d \times H_0^1(\Omega)^d} \tag{11.47}$$

with $f^n$ defined by (11.5) in terms of the external force of the Navier–Stokes model (11.1). We recall that a piecewise linear interpolation in time of $(y^n)_{n\geq 0}$ weakly converges in $L^2(0, T, \boldsymbol{V})$ toward a solution of (11.2)

As done in [2], we may use the least-squares method (analyzed in Section 11.2) to solve iteratively (11.47). Precisely, to approximate $y^{n+1}$ from $y^n$, we may consider the following extremal problem:

$$\inf_{y\in\boldsymbol{V}} E_n(y), \quad E_n(y) = \frac{1}{2}|||v|||_{\boldsymbol{V}}^2, \tag{11.48}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

where the corrector $v \in \boldsymbol{V}$ solves

$$\begin{aligned}\alpha \int_\Omega v \cdot w + \nu \int_\Omega \nabla v \cdot \nabla w = -\alpha \int_\Omega y \cdot w - \nu \int_\Omega \nabla y \cdot \nabla w - \int_\Omega y \cdot \nabla y \cdot w \\ + \langle f^n, w \rangle_{H^{-1}(\Omega)^2 \times H_0^1(\Omega)^2} + \alpha \int_\Omega y^n \cdot w \quad \forall w \in \boldsymbol{V}\end{aligned} \tag{11.49}$$

with $\alpha$ and $f^n$ given by (11.5). For any $n \geq 0$, a minimizing sequence $\{y_k^n\}_{(k\geq 0)}$ for $E_n$ is defined as follows:

$$\begin{cases} y_0^{n+1} = y^n, \\ y_{k+1}^{n+1} = y_k^{n+1} - \lambda_k Y_{1,k}^{n+1}, \quad k \geq 0, \\ \lambda_k = \underset{\lambda \in [0,m]}{\operatorname{argmin}} E_n(y_k^n - \lambda Y_{1,k}^n), \end{cases} \tag{11.50}$$

where $Y_{1,k}^n \in \boldsymbol{V}$ solves (11.30). Remark that, in view of Theorem 11.2.15, the first element of the minimizing sequence is chosen equal to $y^n$, i. e., the minimizer of $E_{n-1}$.

The main goal of this section is to prove that for all $n \in \mathbb{N}$, the minimizing sequence $(y_k^{n+1})_{k\in\mathbb{N}}$ converges to a solution $y^{n+1}$ of (11.47). This allows us to justify the use of least-squares method to solve the backward Euler scheme. Arguing as in Lemma 11.2.14, we have to prove the existence of a constant $c_1 \in (0,1)$ such that $\tau(y_k^n) \leq c_1$ for all $n$ and $k$ in $\mathbb{N}$. Remark that the initialization $y_0^{n+1}$ is fixed as the minimizer of the functional $E^{n-1}$, obtained at the previous iterate. Consequently, the uniform property $\tau(y_k^n) \leq c_1$ is related to the initial guess $y_0^0$ equal to the initial position $u_0$, to the external force $f$ (see (11.2)), and to the value of $\alpha$, where $u_0$ and $f$ are given a priori. On the other hand, the parameter $\alpha$, related to the discretization parameter $\delta t$, can be chosen as large as necessary. As we will see, this uniform property, which is essential to set up the least-squares procedure, requires smallness properties on $u_0$ and $f$.

We start with the following result analogue to Proposition 11.2.3.

**Proposition 11.2.20.** *Let $(f^n)_{n\in\mathbb{N}}$ be a sequence in $H^{-1}(\Omega)^2$, let $\alpha > 0$, and let $y^0 = u_0 \in \boldsymbol{H}$. For any $n \in \mathbb{N}$, there exists a solution $y^{n+1} \in \boldsymbol{V}$ of*

$$\alpha \int_\Omega (y^{n+1} - y^n) \cdot w + \nu \int_\Omega \nabla y^{n+1} \cdot \nabla w + \int_\Omega y^{n+1} \cdot \nabla y^{n+1} \cdot w = \langle f^n, w \rangle_{H^{-1}(\Omega)^2 \times H_0^1(\Omega)^2} \tag{11.51}$$

*for all $w \in \boldsymbol{V}$. Moreover, for all $n \in \mathbb{N}$, $y^{n+1}$ satisfies*

$$\|\!| y^{n+1} |\!\|_{\boldsymbol{V}}^2 \leq \frac{c_0}{\nu} \|f^n\|_{H^{-1}(\Omega)^2}^2 + \alpha \|y^n\|_2^2, \tag{11.52}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*where $c_0 > 0$, only connected to the Poincaré constant, depends on $\Omega$. Moreover, for all $n \in \mathbb{N}^\star$,*

$$\|y^n\|_2^2 + \frac{\nu}{\alpha}\sum_{k=1}^{n}\|\nabla y^k\|_2^2 \leq \frac{1}{\nu}\left(\frac{c_0}{\alpha}\sum_{k=0}^{n-1}\|f^k\|_{H^{-1}(\Omega)^2}^2 + \nu\|u_0\|_2^2\right). \tag{11.53}$$

*Proof.* The existence of $y^{n+1}$ is given in Proposition 11.2.3. Inequality (11.53) is obtained by summing (11.52). □

**Remark 11.2.21.** Arguing as in Lemma 11.2.4, if there exists a solution $y^{n+1}$ in $\boldsymbol{V}$ of (11.49) satisfying $\tau(y^{n+1}) < 1$, then such a solution is unique. In view of Proposition 11.2.20, this holds if notably the quantity

$$\mathcal{M}(f,\alpha,\nu) = \frac{1}{\nu^2}\left(\frac{c_0}{\alpha}\sum_{k=0}^{n-1}\|f^k\|_{H^{-1}(\Omega)^2}^2 + \nu\|u_0\|_2^2\right) \tag{11.54}$$

is small enough.

#### 11.2.4.1 Uniform convergence of the least-squares method with respect to *n*

We have the following convergence for weak solutions of (11.51).

**Theorem 11.2.22.** *Suppose $f \in L^2(0,T;H^{-1}(\Omega)^2)$, $u_0 \in \boldsymbol{V}$, and*

$$c(u_0,f) := \max\left(\frac{1}{\alpha}\|u_0\|_{\boldsymbol{V}}^2\left(\nu + \|u_0\|_2\right)^2 + c_0\|f\|_{L^2(0,T;H^{-1}(\Omega)^2)}^2,\right.$$
$$\left. 2c_0\|f\|_{L^2(0,T;H^{-1}(\Omega)^2)}^2 + \nu\|u_0\|_2^2\right).$$

*Let $\alpha$ be large enough, let $f^n$ be given by* (11.5) *for all $n \in \{0,\dots,N-1\}$, and let $\{y^n\}_{n\in\mathbb{N}}$ in $\boldsymbol{V}$ be a solution of* (11.51). *If there exists a constant $c > 0$ such that*

$$c(u_0,f) \leq c\nu^4, \tag{11.55}$$

*then, for any $n \geq 0$, the minimizing sequence $\{y_k^{n+1}\}_{k\in\mathbb{N}}$ defined by* (11.50) *strongly converges to the unique of solution of* (11.51).

*Proof.* According to Proposition 11.2.13, we have to prove the existence of a constant $c_1 \in (0,1)$ such that, for all $n \in \{0,\dots,N-1\}$ and all $k \in \mathbb{N}$, $\tau(y_k^n) \leq c_1$.

For $n = 0$, as in the previous section, it suffices to takes $\alpha$ large enough to ensure conditions (11.42) with $g = y_0^0 = u_0$, leading to the property $\tau(y_k^0) < c_1$ for all $k \in \mathbb{N}$ and therefore $\tau(y^1) < c_1$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

For the next minimizing sequences, let us recall (see Lemma 11.2.14) that for all $n \in \{0, \dots, N-1\}$ and all $k \in \mathbb{N}$,

$$\|y_k^{n+1}\|_{\boldsymbol{V}} \le \|y^n\|_{\boldsymbol{V}} + \frac{m\sqrt{2}}{\nu(1-c_1)} \frac{\sqrt{E_n(y^n)}}{1 - p_{n,0}(\tilde{\lambda}_{n,0})},$$

where $p_{n,0}(\tilde{\lambda}_{n,0})$ is defined as in the proof of Proposition 11.2.7.

First, since $\|f^n\|^2_{H^{-1}(\Omega)^2} \le \alpha \|f\|^2_{L^2(0,T;H^{-1}(\Omega)^2)}$ for all $n \in \{0, \dots, N-1\}$, we can write

$$\begin{aligned} E_0(y^0) = E_0(u_0) &\le \frac{1}{2\nu}\big(\|u_0\|_{\boldsymbol{V}}(\nu + \|u_0\|_2) + \sqrt{c_0}\|f^0\|_{H^{-1}(\Omega)^2}\big)^2 \\ &\le \frac{1}{\nu}\big(\|u_0\|^2_{\boldsymbol{V}}(\nu + \|u_0\|_2)^2 + c_0\|u_0\|^2_{H^{-1}(\Omega)^2}\big) \\ &\le \frac{\alpha}{\nu}\left(\frac{1}{\alpha}\|u_0\|^2_{\boldsymbol{V}}(\nu + \|u_0\|_2)^2 + c_0\|f\|^2_{L^2(0,T;H^{-1}(\Omega)^2)}\right). \end{aligned}$$

Since $y^n$ is solution of (11.51), it follows from (11.49) that for all $n \in \{1, \dots, N-1\}$,

$$\begin{aligned} E_n(y^n) &\le \frac{c_0}{2\nu}\|f^n - f^{n-1}\|^2_{H^{-1}(\Omega)^2} + \frac{\alpha}{2}\|y^n - y^{n-1}\|_2^2 \\ &\le \frac{\alpha}{\nu}(2c_0\|f\|^2_{L^2(0,T;H^{-1}(\Omega)^2)} + \nu\|u_0\|_2^2). \end{aligned}$$

Therefore $E_n(y^n) \le \frac{\alpha}{\nu} c(u_0, f)$ for all $n \in \{0, \dots, N-1\}$.

Let $c_1 \in (0,1)$ and suppose that $c(u_0, f) < (1-c_1)^4\nu^3$. Then, for any $K \in (0,1)$, there exists $\alpha_0 > 0$ such that $c_{\alpha,\nu}\sqrt{E_n(y^n)} \le K < 1$ for all $\alpha \ge \alpha_0$. We therefore have (see Lemma 11.2.14) that for all $\alpha \ge \alpha_0$, $n \in \{0, \dots, N-1\}$, and $k \in \mathbb{N}$,

$$\begin{aligned} \|y_k^{n+1}\|_{\boldsymbol{V}} &\le \|y^n\|_{\boldsymbol{V}} + \frac{m\sqrt{2}}{\nu(1-c_1)} \frac{\sqrt{E_n(y^n)}}{1 - c_{\alpha,\nu}\sqrt{E_n(y^n)}} \\ &\le \|y^n\|_{\boldsymbol{V}} + \frac{m\sqrt{2}}{\nu(1-c_1)} \frac{\sqrt{E_n(y^n)}}{1-K} \\ &\le \|y^n\|_{\boldsymbol{V}} + \frac{m\sqrt{2\alpha}}{\nu^{3/2}(1-c_1)(1-K)} \sqrt{c(u_0,f)}. \end{aligned} \tag{11.56}$$

Then from (11.53) we obtain, for all $n \in \{0, \dots, N-1\}$,

$$\|y^n\|_{\boldsymbol{V}} \le \frac{\sqrt{\alpha}}{\nu} \sqrt{\frac{c_0}{\alpha} \sum_{k=0}^{n-1} \|f^k\|^2_{H^{-1}(\Omega)^2} + \nu\|u_0\|_2^2},$$

and since $\frac{c_0}{\alpha}\sum_{k=0}^{n-1}\|f^k\|^2_{H^{-1}(\Omega)^2} \le c_0\|f\|^2_{L^2(0,T;H^{-1}(\Omega)^2)}$, we deduce that if

$$c_0\|f\|^2_{L^2(0,T;H^{-1}(\Omega)^2)} + \nu\|u_0\|_2^2 \le \frac{c_1^2}{2}\nu^3,$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

then $\|y^n\|_{\boldsymbol{V}} \leq \frac{c_1}{2}\sqrt{2\alpha\nu}$. Moreover, assuming that $c(u_0,f) \leq \frac{c_1^2(1-c_1)^2(1-K)^2}{4m}\nu^4$, we deduce from (11.56) that for all $n \in \{0,\dots,N-1\}$ and $k \in \mathbb{N}$,

$$\|y_k^n\|_{\boldsymbol{V}} \leq \frac{c_1}{2}\sqrt{2\alpha\nu} + \frac{c_1}{2}\sqrt{2\alpha\nu} = c_1\sqrt{2\alpha\nu},$$

that is, $\tau(y_k^n) \leq c_1$. The result follows from Proposition 11.2.13. □

We emphasize that, for each $n \in \mathbb{N}$, the limit $y^{n+1}$ of the sequence $(y_k^{n+1})_{k\in\mathbb{N}}$ satisfies $\tau(y^{n+1}) < 1$ and is therefore a unique solution of (11.51). Moreover, for $\alpha$ large enough, condition (11.55) reads as the following smallness property on the data $u_0$ and $f$: $c_0\|f\|^2_{L^2(0,T;H^{-1}(\Omega)^2)} + \nu\|u_0\|_2^2 \leq c\nu^4$. In contrast with the static case of Section 11.2, where the unique condition (11.42) on the data $g$ is fulfilled as soon as $\alpha$ is large, the iterated case requires a condition on the data $u_0$ and $f$, whatever be the amplitude of $\alpha$. Again, this smallness property is introduced to guarantee the condition $\tau(y^n) < 1$ for all $n$. In view of (11.53), this condition implies notably that $\|y^n\|_2 \leq c\,\nu^{3/2}$ for all $n > 0$.

For regular solutions of (11.51) we now consider, we may slightly improve the results, notably based on the control of two consecutive elements of the corresponding sequence $(y^n)_{n\in\mathbb{N}}$ for the $L^2$ norm. We first start with the following result of regularity.

**Proposition 11.2.23.** *Assume that $\Omega$ is $C^2$, that $(f^n)_n$ is a sequence in $L^2(\Omega)^2$, and that $u_0 \in \boldsymbol{V}$. Then, for all $n \in \mathbb{N}$, any solution $y^{n+1} \in \boldsymbol{V}$ of* (11.51) *belongs to $H^2(\Omega)^2$.*

*If, moreover, there exists $C > 0$ such that*

$$\frac{c_0}{\alpha}\sum_{k=0}^{n}\|f^k\|^2_{H^{-1}(\Omega)^2} + \nu\|y^0\|_2^2 < C\nu^3, \tag{11.57}$$

*then $y^{n+1}$ satisfies*

$$\int_\Omega |\nabla y^{n+1}|^2 + \frac{\nu}{2\alpha}\sum_{k=1}^{n+1}\int_\Omega |P\Delta y^k|^2 \leq \frac{1}{\nu}\left(\frac{1}{\alpha}\sum_{k=0}^{n}\|f^k\|_2^2 + \nu\|\nabla u_0\|_2^2\right), \tag{11.58}$$

*where $P$ is the operator of projection from $L^2(\Omega)^2$ into $\boldsymbol{H}$.*

*Proof.* From Proposition 11.2.3 we know that $y^n \in H^2(\Omega)^2 \cap \boldsymbol{V}$ for all $n \in \mathbb{N}^*$. Thus, integrating by parts (11.51) and using the density argument, we obtain

$$\alpha\int_\Omega (y^{n+1}-y^n)\cdot w - \nu\int_\Omega \Delta y^{n+1} w + \int_\Omega y^{n+1}\cdot\nabla y^{n+1}\cdot w = \langle f^n, w\rangle_{H^{-1}(\Omega)^2\times H_0^1(\Omega)^2} \tag{11.59}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

for all $w \in \boldsymbol{H}$. Then taking $w = P\Delta y^{n+1}$ and integrating by part lead to

$$\begin{aligned}\alpha \int_\Omega |\nabla y^{n+1}|^2 + \nu \int_\Omega |P\Delta y^{n+1}|^2 = &- \int_\Omega f^n P\Delta y^{n+1} + \int_\Omega y^{n+1} \cdot \nabla y^{n+1} \cdot P\Delta y^{n+1} \\ &+ \alpha \int_\Omega \nabla y^n \cdot \nabla y^{n+1}.\end{aligned} \tag{11.60}$$

Recall that

$$\int_\Omega f^n P\Delta y^{n+1} \le \frac{1}{2\nu}\|f^n\|_2^2 + \frac{\nu}{2}\|P\Delta y^{n+1}\|_2^2, \quad \alpha \int_\Omega \nabla y^n \cdot \nabla y^{n+1} \le \frac{\alpha}{2}\|y^n\|_{\boldsymbol{V}}^2 + \frac{\alpha}{2}\|y^{n+1}\|_{\boldsymbol{V}}^2.$$

We also have

$$\left| \int_\Omega y^{n+1} \cdot \nabla y^{n+1} \cdot P\Delta y^{n+1} \right| \le \|y^{n+1}\|_\infty \|\nabla y^{n+1}\|_2 \|P\Delta y^{n+1}\|_2.$$

We now use that there exist three constants $c_1$, $c_2$, and $c_3$ such that

$$\|\Delta y^{n+1}\|_2 \le c_1 \|P\Delta y^{n+1}\|_2, \quad \|y^{n+1}\|_\infty \le c_2 \|y^{n+1}\|_2^{\frac{1}{2}} \|\Delta y^{n+1}\|_2^{\frac{1}{2}}$$

and

$$\|\nabla y^{n+1}\|_2 \le c_3 \|y^{n+1}\|_2^{\frac{1}{2}} \|\Delta y^{n+1}\|_2^{\frac{1}{2}}$$

(for the second inequality, see [24, Chapter 25, p. 144]). This implies that (for $c = c_1 c_2 c_3$)

$$\left| \int_\Omega y^{n+1} \cdot \nabla y^{n+1} \cdot P\Delta y^{n+1} \right| \le c \|y^{n+1}\|_2 \|P\Delta y^{n+1}\|_2^2.$$

Recalling (11.60), it follows that

$$\frac{\alpha}{2} \int_\Omega |\nabla y^{n+1}|^2 + \left( \frac{\nu}{2} - c\|y^{n+1}\|_2 \right) \int_\Omega |P\Delta y^{n+1}|^2 \le \frac{1}{2\nu}\|f^n\|_2^2 + \frac{\alpha}{2} \int_\Omega |\nabla y^n|^2.$$

By estimate (11.53) assumption (11.57) implies that $\|y^{n+1}\|_2 \le \frac{\nu}{4c}$ and

$$\int_\Omega |\nabla y^{n+1}|^2 + \frac{\nu}{2\alpha} \int_\Omega |P\Delta y^{n+1}|^2 \le \frac{1}{\nu\alpha}\|f^n\|_2^2 + \int_\Omega |\nabla y^n|^2.$$

Summing then implies (11.58) for all $n \in \mathbb{N}$. □

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Remark 11.2.24.** Under the hypotheses of Proposition 11.2.23, suppose that

$$B_{\alpha,\nu} := (\alpha\nu^5)^{-1}\left(c_0\alpha^{-1}\sum_{k=0}^{n}\|f^k\|^2_{H^{-1}(\Omega)^2} + \nu\|y^0\|_2^2\right)\left(\alpha^{-1}\sum_{k=0}^{n-1}\|f^k\|_2^2 + \nu\|\nabla y^0\|_2^2\right)$$

is small (which is satisfied as soon as $\alpha$ is large enough). Then the solution of (11.51) is unique.

Indeed, let $n \in \mathbb{N}$, and let $y_1^{n+1}, y_2^{n+1} \in \boldsymbol{V}$ be two solutions of (11.51). Then $Y := y_1^{n+1} - y_2^{n+1}$ satisfies

$$\alpha\int_\Omega Y\cdot w + \nu\int_\Omega \nabla Y\cdot\nabla w + \int_\Omega y_2^{n+1}\cdot\nabla Y\cdot w + \int_\Omega Y\cdot\nabla y_1^{n+1}\cdot w = 0 \quad \forall w\in\boldsymbol{V},$$

and, in particular, for $w = Y$ (since $\int_\Omega y_2^{n+1}\cdot\nabla Y\cdot Y = 0$),

$$\begin{aligned}\alpha\int_\Omega |Y|^2 + \nu\int_\Omega|\nabla Y|^2 = -\int_\Omega Y\cdot\nabla y_1^{n+1}\cdot Y &= \int_\Omega Y\cdot\nabla Y\cdot y_1^{n+1}\\ &\le c\|y_1^{n+1}\|_\infty\|\nabla Y\|_2\|Y\|_2\\ &\le c\|y_1^{n+1}\|_2^{1/2}\|P\Delta y_1^{n+1}\|_2^{1/2}\|\nabla Y\|_2\|Y\|_2\\ &\le \alpha\|Y\|_2^2 + \frac{c}{\alpha}\|y_1^{n+1}\|_2\|P\Delta y_1^{n+1}\|_2\|\nabla Y\|_2^2,\end{aligned}$$

leading to

$$\left(\nu - \frac{c}{\alpha}\|y_1^{n+1}\|_2\|P\Delta y_1^{n+1}\|_2\right)\|\nabla Y\|_2^2 \le 0.$$

If

$$\|y_1^{n+1}\|_2\|P\Delta y_1^{n+1}\|_2 < \frac{\nu\alpha}{c}, \tag{11.61}$$

then $Y = 0$, and the solution is unique. However, by (11.53) and (11.58) we have

$$\|y_1^{n+1}\|_2^2\|P\Delta y_1^{n+1}\|_2^2 \le \frac{4\alpha}{\nu^3}\left(\frac{c_0}{\alpha}\sum_{k=0}^{n}\|f^k\|^2_{H^{-1}(\Omega)^2} + \nu\|y^0\|_2^2\right)\left(\frac{1}{\alpha}\sum_{k=0}^{n}\|f^k\|_2^2 + \nu\|\nabla y^0\|_2^2\right).$$

Therefore, if there exists a constant $C$ such that $B_{\alpha,\nu} < C$, then (11.61) holds.

Proposition 11.2.23 then allows us to obtain the following estimate of $\|y^{n+1} - y^n\|_2$ in terms of the parameter $\alpha$.

**Theorem 11.2.25.** *We assume that $\Omega$ is $C^2$, that $(f^n)_n$ is a sequence in $L^2(\Omega)^2$ satisfies $\alpha^{-1}\sum_{k=0}^{+\infty}\|f^k\|_2^2 < +\infty$, that $u_0 \in \boldsymbol{V}$, and that for all $n\in\mathbb{N}$, $y^{n+1}\in H^2(\Omega)^2\cap\boldsymbol{V}$ is a solution*

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*of* (11.51) *satisfying* $\|y^{n+1}\|_2 \leq \frac{\nu}{4c}$. *Then there exists* $C_1 > 0$ *such that the sequence* $(y^n)_n$ *satisfies*

$$\|y^{n+1} - y^n\|_2^2 \leq \frac{C_1}{\alpha\nu^{3/2}}. \tag{11.62}$$

*Proof.* For all $n \in \mathbb{N}$, $w = y^{n+1} - y^n$ in (11.51) gives

$$\alpha\|y^{n+1} - y^n\|_2^2 + \nu\|\nabla y^{n+1}\|_2^2 \leq \left|\int_\Omega y^{n+1}.\nabla y^{n+1}.(y^{n+1} - y^n)\right| + \left|\int_\Omega f^n.(y^{n+1} - y^n)\right| + \nu\left|\int_\Omega \nabla y^n.\nabla y^{n+1}\right|.$$

Moreover,

$$\left|\int_\Omega y^{n+1}.\nabla y^{n+1}.(y^{n+1} - y^n)\right| \leq c\|\nabla y^{n+1}\|_2^2\|\nabla(y^{n+1} - y^n)\|_2 \leq c\|\nabla y^{n+1}\|_2^2(\|\nabla y^{n+1}\|_2 + \|\nabla y^n\|_2).$$

Therefore

$$\alpha\|y^{n+1} - y^n\|_2^2 + \nu\|\nabla y^{n+1}\|_2^2 \leq c\|\nabla y^{n+1}\|_2^2(\|\nabla y^{n+1}\|_2 + \|\nabla y^n\|_2) + \frac{1}{\alpha}\|f^n\|_2^2 + \nu\|\nabla y^n\|_2^2.$$

However, (11.58) implies that for all $n \in \mathbb{N}$,

$$\int_\Omega |\nabla y^{n+1}|^2 \leq \frac{1}{\nu}\left(\frac{1}{\alpha}\sum_{k=0}^{+\infty}\|f^k\|_2^2 + \nu\|\nabla y^0\|_2^2\right) := \frac{C}{\nu},$$

and thus, since $\nu < 1$,

$$\alpha\|y^{n+1} - y^n\|_2^2 + \nu\|\nabla y^{n+1}\|_2^2 \leq \frac{2cC^{3/2}}{\nu^{3/2}} + 2C \leq \frac{C_1}{\nu^{3/2}},$$

leading to $\|y^{n+1} - y^n\|_2^2 = O(\frac{1}{\alpha\nu^{3/2}})$ as claimed. □

This result states that two consecutive elements of the sequence $(y^n)_{n\geq 0}$ defined by recurrence from the scheme (11.3) are close to each other as soon as the discretization time step $\delta t$ is small enough. In particular, this justifies the choice of the initial term $y_0^{n+1} = y^n$ of the minimizing sequence to approximate $y^{n+1}$.

We end this section with the following result, analogous to Theorem 11.2.22, for regular data.

**Theorem 11.2.26.** *Suppose* $f \in L^2(0,T;L^2(\Omega)^2)$, $u_0 \in \boldsymbol{V}$, *for all* $n \in \{0,\dots,N-1\}$, $\alpha$ *and* $f^n$ *are given by* (11.5), *and* $y^{n+1} \in \boldsymbol{V}$ *is a solution of* (11.51). *If* $C(u_0,f) := \|f\|^2_{L^2(0,T;L^2(\Omega)^2)} +$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$\nu\|u_0\|_{\boldsymbol{V}}^2 \leq Cv^2$ *for some* $C$ *and* $\alpha$ *is large enough, then, for any* $n \geq 0$, *the minimizing sequence* $\{y_k^{n+1}\}_{k\in\mathbb{N}}$ *defined by* (11.50) *strongly converges to the unique of solution of* (11.51).

*Proof.* As for Theorem 11.2.22, it suffices to prove that there exists $c_1 \in (0,1)$ such that $\tau(y_k^n) \leq c_1$ for all $n \in \{0,\dots,N-1\}$ and $k \in \mathbb{N}$. Let us recall that for all $n \in \{0,\dots,N-1\}$ and $k \in \mathbb{N}$,

$$\|y_{k+1}^{n+1}\|_{\boldsymbol{V}} \leq \|y^n\|_{\boldsymbol{V}} + \frac{m\sqrt{2}}{\nu(1-c_1)}\frac{\sqrt{E_n(y^n)}}{1-p_{n,0}(\widetilde{\lambda}_{n,0})},$$

where $p_{n,0}(\widetilde{\lambda}_{n,0})$ is defined as in the proof of Proposition 11.2.7. From (11.49), since $\|f^n\|_2^2 \leq \alpha\|f\|_{L^2(0,T;L^2(\Omega)^2)}^2$ for all $n \in \{0,\dots,N-1\}$, we have

$$\begin{aligned} E_0(y^0) = E_0(u_0) &\leq \frac{1}{2\nu}\left(\|u_0\|_{\boldsymbol{V}}(\nu + \|u_0\|_2) + \sqrt{\frac{\nu}{\alpha}}\|f^1\|_2\right)^2 \\ &\leq \frac{1}{\nu}\|u_0\|_{\boldsymbol{V}}^2(\nu + \|u_0\|_2)^2 + \|f\|_{L^2(0,T;L^2(\Omega)^2),}^2 \end{aligned}$$

and since $y^n$ is a solution of (11.51), for all $n \in \{1,\dots,N-1\}$, we have

$$\begin{aligned} E_n(y^n) &\leq \frac{1}{\alpha}\|f^n - f^{n-1}\|_2^2 + \alpha\|y^n - y^{n-1}\|_2^2 \\ &\leq 2\|f\|_{L^2(0,T;L^2(\Omega)^2)}^2 + \alpha\|y^n - y^{n-1}\|_2^2. \end{aligned}$$

From the proof of Theorem 11.2.25 we deduce that for all $n \in \{0,\dots,N-1\}$,

$$\alpha\|y^{n+1} - y^n\|_2^2 \leq \frac{2cC(u_0,f)^{3/2}}{\nu^{3/2}} + 2C(u_0,f),$$

and thus, for all $n \in \{1,\dots,N-1\}$,

$$E_n(y^n) \leq \frac{2cC(u_0,f)^{3/2}}{\nu^{3/2}} + 4C(u_0,f).$$

Moreover, from (11.58), for all $n \in \{0,\dots,N-1\}$,

$$\|y^n\|_{\boldsymbol{V}}^2 \leq \frac{1}{\nu}\left(\frac{1}{\alpha}\sum_{k=0}^{n}\|f^k\|_2^2 + \nu\|u_0\|_{\boldsymbol{V}}^2\right) \leq \frac{1}{\nu}(\|f\|_{L^2(0,T;L^2(\Omega)^2)}^2 + \nu\|u_0\|_{\boldsymbol{V}}^2) = \frac{1}{\nu}C(u_0,f).$$

Eventually, let $c_1 \in (0,1)$. Then there exists $\alpha_0 > 0$ such that $c_{\alpha,\nu}\sqrt{E_n(y^n)} \leq K < 1$ for all $\alpha \geq \alpha_0$. We therefore have (see Theorem 11.2.22) that for all $\alpha \geq \alpha_0$, all $n \in \{0,\dots,N-1\}$, and all $k \in \mathbb{N}$,

$$\|y_{k+1}^{n+1}\|_{\boldsymbol{V}} \leq \|y^n\|_{\boldsymbol{V}} + \frac{m\sqrt{2}}{\nu(1-c_1)}\frac{\sqrt{E_n(y^n)}}{1-K},$$

which gives a bound of $\|y_{k+1}^{n+1}\|_{\boldsymbol{V}}$ independent of $\alpha \geq \alpha_0$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Taking $\alpha_1 \geq \alpha_0$ large enough, we deduce that $\|y_k^n\|_{\boldsymbol{V}} \leq c_1\sqrt{2\alpha\nu}$ for all $\alpha \geq \alpha_1$, $n \in \{0,\dots,N-1\}$, and $k \in \mathbb{N}$, that is, $\tau_2(y_k^n) \leq c_1$. The announced convergence follows from Proposition 11.2.13. □

## 11.3 Space-time least squares method

Adapting the previous section, we introduce and analyze a least-squares functional allowing us to approximate the solution of the boundary value problem (11.2). Though more technical, the analysis is simpler since (11.2) (in this 2D setting) admits a unique weak solution, independently of the size of the data, contrary to (11.4). Consequently, we will see that this space-time setting alleviates the smallness assumptions on the data $u_0$ and $f$.

### 11.3.1 Preliminary technical results

In the following, we repeatedly use the following classical estimate.

**Lemma 11.3.1.** *Let $u \in \boldsymbol{H}$ and $v, w \in \boldsymbol{V}$. There exists a constant $c = c(\Omega)$ such that*

$$\int_\Omega u \cdot \nabla v \cdot w \leq c\|u\|_{\boldsymbol{H}}\|v\|_{\boldsymbol{V}}\|w\|_{\boldsymbol{V}}. \tag{11.63}$$

*Proof.* If $u \in \boldsymbol{H}$ and $v, w \in \boldsymbol{V}$, then denoting $\tilde{u}$, $\tilde{v}$, and $\tilde{w}$ their extensions to 0 in $\mathbb{R}^2$, we have $\tilde{u} \cdot \nabla\tilde{v} \in \mathcal{H}^1(\mathbb{R}^2)^2$ and $\|\tilde{u} \cdot \nabla\tilde{v}\|_{\mathcal{H}^1(\mathbb{R}^2)^2} \leq \|\tilde{u}\|_2\|\nabla\tilde{v}\|_2$ (see [3, Theorem II.1 and Remark II.2]). Moreover, $BMO(\mathbb{R}^2) = (\mathcal{H}^1(\mathbb{R}^2))'$ and $H^1(\mathbf{R}^2) \subset BMO(\mathbb{R}^2)$ (see [24, Chapter 4, p. 21, and Chapter 16, p. 88]), and thus

$$\begin{aligned}\left|\int_\Omega u \cdot \nabla v \cdot w\right| = \left|\int_\Omega \tilde{u} \cdot \nabla\tilde{v} \cdot \tilde{w}\right| &\leq \|\tilde{u} \cdot \nabla\tilde{v}\|_{\mathcal{H}^1(\mathbb{R}^2)^2}\|\tilde{w}\|_{BMO(\mathbb{R}^2)^2} \\ &\leq c\|\tilde{u}\|_2\|\nabla\tilde{v}\|_2\|\tilde{w}\|_{H^1(\mathbb{R}^2)^2} \\ &\leq c\|u\|_2\|\nabla v\|_2\|w\|_{H^1(\Omega)^2} \leq c\|u\|_{\boldsymbol{H}}\|v\|_{\boldsymbol{V}}\|w\|_{\boldsymbol{V}}.\end{aligned}$$

□

**Lemma 11.3.2.** *Let $u \in L^\infty(0,T;\boldsymbol{H})$ and $v \in L^2(0,T;\boldsymbol{V})$. Then the function $B(u,v)$ defined by*

$$\langle B(u(t),v(t)),w\rangle = \int_\Omega u(t) \cdot \nabla v(t) \cdot w \quad \forall w \in \boldsymbol{V}, \text{ for a. e. } t \in [0,T]$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*belongs to* $L^2(0,T;\boldsymbol{V}')$, *and*

$$\|B(u,v)\|_{L^2(0,T;\boldsymbol{V}')} \leq c\left(\int_0^T \|u\|_{\boldsymbol{H}}^2 \|v\|_{\boldsymbol{V}}^2\right)^{\frac{1}{2}} \leq c\|u\|_{L^\infty(0,T;\boldsymbol{H})}\|v\|_{L^2(0,T;\boldsymbol{V})}. \tag{11.64}$$

*Moreover,*

$$\langle B(u,v),v\rangle_{\boldsymbol{V}'\times\boldsymbol{V}} = 0. \tag{11.65}$$

*Proof.* Indeed, for a. e. $t \in [0,T]$, we have (see (11.63))

$$\left|\langle B(u(t),v(t)),w\rangle\right| \leq c\|u(t)\|_{\boldsymbol{H}}\|v(t)\|_{\boldsymbol{V}}\|w\|_{\boldsymbol{V}} \quad \forall w \in \boldsymbol{V},$$

and thus

$$\int_0^T \|B(u,v)\|_{\boldsymbol{V}'}^2 \leq c\int_0^T \|u\|_{\boldsymbol{H}}^2\|v\|_{\boldsymbol{V}}^2 \leq c\|u\|_{L^\infty(0,T;\boldsymbol{H})}^2\|v\|_{L^2(0,T,\boldsymbol{V})}^2 < +\infty.$$

We also have that for a. e. $t \in [0,T]$ (see [25]),

$$\langle B(u(t),v(t)),v(t)\rangle_{\boldsymbol{V}'\times\boldsymbol{V}} = \int_\Omega u(t)\cdot\nabla v(t)\cdot v(t) = 0.$$

□

**Lemma 11.3.3.** *Let* $u \in L^2(0,T;\boldsymbol{V})$. *Then the function* $B_1(u)$ *defined by*

$$\langle B_1(u(t)),w\rangle = \int_\Omega \nabla u(t)\cdot\nabla w \quad \forall w \in \boldsymbol{V},\ \textit{for a. e. } t \in [0,T],$$

*belongs to* $L^2(0,T;\boldsymbol{V}')$, *and*

$$\|B_1(u)\|_{L^2(0,T;\boldsymbol{V}')} \leq \|u\|_{L^2(0,T;\boldsymbol{V})} < +\infty. \tag{11.66}$$

*Proof.* Indeed, for a.e, $t \in [0,T]$, we have

$$\left|\langle B_1(u(t)),w\rangle\right| \leq \|\nabla u(t)\|_2\|\nabla w\|_2 = \|u(t)\|_{\boldsymbol{V}}\|w\|_{\boldsymbol{V}},$$

and thus for a. e. $t \in [0,T]$,

$$\|B_1(u(t))\|_{\boldsymbol{V}'} \leq \|u(t)\|_{\boldsymbol{V}},$$

which gives (11.66). □

We also have (see [16, 25]) the following:

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Lemma 11.3.4.** *For all $y \in L^2(0,T;\boldsymbol{V}) \cap H^1(0,T;\boldsymbol{V}')$, we have $y \in \mathcal{C}([0,T];\boldsymbol{H})$ and in $\mathcal{D}'(0,T)$, for all $w \in \boldsymbol{V}$,*

$$\langle \partial_t y, w\rangle_{\boldsymbol{V}'\times\boldsymbol{V}} = \int_\Omega \partial_t y \cdot w = \frac{d}{dt}\int_\Omega y \cdot w, \quad \langle \partial_t y, y\rangle_{\boldsymbol{V}'\times\boldsymbol{V}} = \frac{1}{2}\frac{d}{dt}\int_\Omega |y|^2 \tag{11.67}$$

*and*

$$\|y\|^2_{L^\infty(0,T;\boldsymbol{H})} \le c\|y\|_{L^2(0,T;\boldsymbol{V})}\|\partial_t y\|_{L^2(0,T;\boldsymbol{V}')}. \tag{11.68}$$

We recall that along this section, we suppose that $u_0 \in \boldsymbol{H}$, $f \in L^2(0,T;\boldsymbol{V}')$ and $\Omega$ is a bounded Lipschitz domain of $\mathbb{R}^2$. We also denote

$$\mathcal{A} = \{y \in L^2(0,T;\boldsymbol{V}) \cap H^1(0,T;\boldsymbol{V}'),\ y(0) = u_0\}$$

and

$$\mathcal{A}_0 = \{y \in L^2(0,T;\boldsymbol{V}) \cap H^1(0,T;\boldsymbol{V}'),\ y(0) = 0\}.$$

Endowed with the scalar product

$$\langle y, z\rangle_{\mathcal{A}_0} = \int_0^T \langle y, z\rangle_{\boldsymbol{V}} + \langle \partial_t y, \partial_t z\rangle_{\boldsymbol{V}'}$$

and the associated norm

$$\|y\|_{\mathcal{A}_0} = \sqrt{\|y\|^2_{L^2(0,T;\boldsymbol{V})} + \|\partial_t y\|^2_{L^2(0,T;\boldsymbol{V}')}},$$

$\mathcal{A}_0$ is a Hilbert space.

We also recall and introduce several technical results. The first one is well known (we refer to [16] and [25]).

**Proposition 11.3.5.** *There exists a unique $\bar{y} \in \mathcal{A}$ solution in $\mathcal{D}'(0,T)$ of* (11.2). *This solution satisfies the following estimates:*

$$\|\bar{y}\|^2_{L^\infty(0,T;\boldsymbol{H})} + \nu\|\bar{y}\|^2_{L^2(0,T;\boldsymbol{V})} \le \|u_0\|^2_{\boldsymbol{H}} + \frac{1}{\nu}\|f\|^2_{L^2(0,T;\boldsymbol{V}')},$$
$$\|\partial_t \bar{y}\|_{L^2(0,T;\boldsymbol{V}')} \le \sqrt{\nu}\|u_0\|_{\boldsymbol{H}} + 2\|f\|_{L^2(0,T;\boldsymbol{V}')} + \frac{c}{\nu^{\frac{3}{2}}}\left(\nu\|u_0\|^2_{\boldsymbol{H}} + \|f\|^2_{L^2(0,T;\boldsymbol{V}')}\right).$$

We also introduce the following:

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Proposition 11.3.6.** *For all $y \in L^2(0,T,\boldsymbol{V}) \cap H^1(0,T;\boldsymbol{V}')$, there exists a unique solution $v \in \mathcal{A}_0$ in $\mathcal{D}'(0,T)$ of*

$$\begin{cases} \dfrac{d}{dt}\displaystyle\int_\Omega v\cdot w + \int_\Omega \nabla v\cdot\nabla w + \frac{d}{dt}\int_\Omega y\cdot w + \nu\int_\Omega \nabla y\cdot\nabla w \\ \qquad + \displaystyle\int_\Omega y\cdot\nabla y\cdot w = \langle f, w\rangle_{\boldsymbol{V}'\times\boldsymbol{V}} \quad \forall w\in\boldsymbol{V}, \\ v(0) = 0. \end{cases} \tag{11.69}$$

*Moreover, for all $t \in [0,T]$,*

$$\|v(t)\|_{\boldsymbol{H}}^2 + \|v\|_{L^2(0,t;\boldsymbol{V})}^2 \le \|f - B(y,y) - \nu B_1(y) - \partial_t y\|_{L^2(0,t;\boldsymbol{V}')}^2$$

*and*

$$\begin{aligned} \|\partial_t v\|_{L^2(0,T;\boldsymbol{V}')} &\le \|v\|_{L^2(0,T,\boldsymbol{V})} + \|f - B(y,y) - \nu B_1(y) - \partial_t y\|_{L^2(0,T;\boldsymbol{V}')} \\ &\le 2\|f - B(y,y) - \nu B_1(y) - \partial_t y\|_{L^2(0,T;\boldsymbol{V}')}. \end{aligned}$$

The proof of this proposition is a consequence of the following standard result (see [16, 24]).

**Proposition 11.3.7.** *For all $z_0 \in \boldsymbol{H}$ and $F \in L^2(0,T;\boldsymbol{V}')$, there exists a unique solution $z \in L^2(0,T,\boldsymbol{V}) \cap H^1(0,T;\boldsymbol{V}')$ in $\mathcal{D}'(0,T)$ of*

$$\begin{cases} \dfrac{d}{dt}\displaystyle\int_\Omega z\cdot w + \int_\Omega \nabla z\cdot\nabla w = \langle F, w\rangle_{\boldsymbol{V}'\times\boldsymbol{V}} \quad \forall w\in\boldsymbol{V}, \\ z(0) = z_0. \end{cases} \tag{11.70}$$

*Moreover, for all $t \in [0,T]$,*

$$\|z(t)\|_{\boldsymbol{H}}^2 + \|z\|_{L^2(0,t;\boldsymbol{V})}^2 \le \|F\|_{L^2(0,t;\boldsymbol{V}')}^2 + \|z_0\|_{\boldsymbol{H}}^2 \tag{11.71}$$

*and*

$$\|\partial_t z\|_{L^2(0,T;\boldsymbol{V}')} \le \|z\|_{L^2(0,T,\boldsymbol{V})} + \|F\|_{L^2(0,T;\boldsymbol{V}')} \le 2\|F\|_{L^2(0,T;\boldsymbol{V}')} + \|z_0\|_{\boldsymbol{H}}. \tag{11.72}$$

*Proof of Proposition* 11.3.6. Let $y \in L^2(0,T,\boldsymbol{V}) \cap H^1(0,T;\boldsymbol{V}')$. Then the functions $B(y,y)$ and $B_1(y)$ defined in $\mathcal{D}'(0,T)$ by

$$\langle B(y,y), w\rangle = \int_\Omega y\cdot\nabla y\cdot w \quad \text{and} \quad \langle B_1(y), w\rangle = \int_\Omega \nabla y\cdot\nabla w \quad \forall w\in\boldsymbol{V}$$

belong to $L^2(0,T;\boldsymbol{V}')$ (see Lemmas 11.3.2 and 11.3.3).

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Moreover, since $y \in L^2(0,T,\boldsymbol{V}) \cap H^1(0,T,\boldsymbol{V}')$, in view of (11.67), in $\mathcal{D}'(0,T)$, for all $w \in \boldsymbol{V}$, we have

$$\frac{d}{dt}\int_\Omega y \cdot w = \langle \partial_t y, w\rangle_{\boldsymbol{V}'\times\boldsymbol{V}}.$$

Then (11.9) may be rewritten as

$$\begin{cases} \dfrac{d}{dt}\displaystyle\int_\Omega v\cdot w + \int_\Omega \nabla v\cdot\nabla w = \langle F, w\rangle_{\boldsymbol{V}'\times\boldsymbol{V}} \quad \forall w \in \boldsymbol{V}, \\ v(0) = 0, \end{cases}$$

where $F = f - B(y,y) - \nu B_1(y) - \partial_t y \in L^2(0,T,\boldsymbol{V}')$. Proposition 11.3.6 is therefore a consequence of Proposition 11.3.7. □

## 11.3.2 The least-squares functional

We now introduce our least-squares functional $E : H^1(0,T,\boldsymbol{V}') \cap L^2(0,T,\boldsymbol{V}) \to \mathbb{R}^+$ by putting

$$E(y) = \frac{1}{2}\int_0^T \|v\|_{\boldsymbol{V}}^2 + \frac{1}{2}\int_0^T \|\partial_t v\|_{\boldsymbol{V}'}^2 = \frac{1}{2}\|v\|_{\mathcal{A}_0}^2, \tag{11.73}$$

where the corrector $v$ is a unique solution of (11.69). The infimum of $E$ over the set $\mathcal{A}$ is equal to zero and is reached by a solution of (11.2). In this sense the functional $E$ is a so-called error functional, which measures, through the corrector variable $v$, the deviation of $y$ from being a solution of the underlying equation (11.2).

Beyond this statement, we would like to argue why we believe it is a good idea to use a (minimization) least-squares approach to approximate the solution of (11.2) by minimizing the functional $E$. Our main result of this section is a follows.

**Theorem 11.3.8.** *Let $(y_k)_{k\in\mathbb{N}}$ be a sequence of $\mathcal{A}$ bounded in $L^2(0,T,\boldsymbol{V}) \cap H^1(0,T;\boldsymbol{V}')$. If $E'(y_k) \to 0$ as $k \to \infty$, then the whole sequence $(y_k)_{k\in\mathbb{N}}$ converges strongly as $k \to \infty$ in $L^2(0,T,\boldsymbol{V}) \cap H^1(0,T;\boldsymbol{V}')$ to the solution $\bar{y}$ of* (11.2).

As in the previous section, we divide the proof into two main steps.

1. First, we use a typical a priori bound to show that leading the error functional $E$ down to zero implies the strong convergence to the unique solution of (11.2).
2. Next, we show that taking the derivative $E'$ to zero actually suffices to take zero $E$.

Before to prove this result, we mention the following equivalence, which justifies the least-squares terminology we have used in the following sense: the minimization of

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

the functional $E$ is equivalent to the minimization of the $L^2(0,T,\boldsymbol{V}')$-norm of the main equation of the Navier–Stokes system.

**Lemma 11.3.9.** *There exist $c_1 > 0$ and $c_2 > 0$ such that*

$$c_1 E(y) \leq \left\| y_t + \nu B_1(y) + B(y,y) - f \right\|^2_{L^2(0,T;\boldsymbol{V}')} \leq c_2 E(y)$$

*for all $y \in L^2(0,T,\boldsymbol{V}) \cap H^1(0,T;\boldsymbol{V}')$.*

*Proof.* From Proposition 11.3.6 we deduce that

$$2E(y) = \|v\|^2_{\mathcal{A}_0} \leq 5\left\| y_t + \nu B_1(y) + B(y,y) - f \right\|^2_{L^2(0,T;\boldsymbol{V}')}.$$

On the other hand, by the definition of $v$,

$$\begin{aligned}\left\| y_t + \nu B_1(y) + B(y,y) - f \right\|_{L^2(0,T;\boldsymbol{V}')} &= \left\| v_t + B_1(v) \right\|_{L^2(0,T;\boldsymbol{V}')} \leq \|v_t\|_{L^2(0,T;\boldsymbol{V}')} + \left\| B_1(v) \right\|_{L^2(0,T;\boldsymbol{V}')} \\ &\leq \sqrt{2}\|v\|_{\mathcal{A}_0} = 2\sqrt{E(y)}.\end{aligned}$$

□

We start with the following proposition, which establishes that as we take down the error $E$ to zero, we get closer, in the norm $L^2(0,T;\boldsymbol{V})$ and $H^1(0,T;\boldsymbol{V}')$, to the solution $\bar{y}$ of problem (11.2), and so it justifies why a promising strategy to find good approximations of the solution of problem (11.2) is to look for global minimizers of (11.73).

**Proposition 11.3.10.** *Let $\bar{y} \in \mathcal{A}$ be the solution of* (11.2)*, let $M \in \mathbb{R}$ be such that $\|\partial_t \bar{y}\|_{L^2(0,T,\boldsymbol{V}')} \leq M$ and $\sqrt{\nu}\|\nabla \bar{y}\|_{L^2(Q_T)^4} \leq M$, and let $y \in \mathcal{A}$. If $\|\partial_t y\|_{L^2(0,T,\boldsymbol{V}')} \leq M$ and $\sqrt{\nu}\|\nabla y\|_{L^2(Q_T)^4} \leq M$, then there exists a constant $c(M,\nu)$ such that*

$$\|y - \bar{y}\|_{L^\infty(0,T;\boldsymbol{H})} + \sqrt{\nu}\|y - \bar{y}\|_{L^2(0,T;\boldsymbol{V})} + \|\partial_t y - \partial_t \bar{y}\|_{L^2(0,T,\boldsymbol{V}')} \leq c(M,\nu)\sqrt{E(y)}. \tag{11.74}$$

*Proof.* Let $Y = y - \bar{y}$. The functions $B(Y,y)$, $B(\bar{y},Y)$, and $B_1(v)$ defined in $\mathcal{D}'(0,T)$ by

$$\begin{aligned}\langle B(Y,y), w\rangle &= \int_\Omega Y \cdot \nabla y \cdot w, \quad \langle B(\bar{y},Y), w\rangle = \int_\Omega \bar{y} \cdot \nabla Y \cdot w, \\ \langle B_1(v), w\rangle &= \int_\Omega \nabla v \cdot \nabla w \quad \forall w \in \boldsymbol{V}\end{aligned}$$

belong to $L^2(0,T;\boldsymbol{V}')$ (see Lemmas 11.3.2 and 11.3.3), and from (11.2), (11.69), and (11.67) we deduce that

$$\begin{cases} \dfrac{d}{dt}\displaystyle\int_\Omega Y \cdot w + \nu \int_\Omega \nabla Y \cdot \nabla w = -\langle \partial_t v + B_1(v) + B(Y,y) + B(\bar{y},Y), w\rangle_{\boldsymbol{V}'\times\boldsymbol{V}} \quad \forall w \in \boldsymbol{V}, \\ Y(0) = 0, \end{cases}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

and from (11.71), (11.72), (11.64), (11.65), and (11.66) we deduce that for all $t \in [0,T]$,

$$\begin{aligned}\int_\Omega \big|Y(t)\big|^2 + \nu \int_{Q_t} |\nabla Y|^2 &\le \frac{1}{\nu}\int_0^t \big\|\partial_t v + B_1(v) + B(Y,y)\big\|_{\boldsymbol{V}'}^2 \\ &\le \frac{4}{\nu}\left(\|\partial_t v\|^2_{L^2(0,T,\boldsymbol{V}')} + \|v\|^2_{L^2(0,T,\boldsymbol{V})} + c\int_0^t \|Y\|_2^2\|y\|_{\boldsymbol{V}}^2\right) \\ &\le \frac{4}{\nu}\left(2E(y) + c\int_0^t \|Y\|_2^2\|y\|_{\boldsymbol{V}}^2\right).\end{aligned}$$

Grönwall's lemma then implies that for all $t \in [0,T]$,

$$\int_\Omega \big|Y(t)\big|^2 + \nu \int_{Q_t} |\nabla Y|^2 \le \frac{8}{\nu}E(y)\exp\left(\frac{c}{\nu}\int_0^t \|y\|_{\boldsymbol{V}}^2\right) \le \frac{8}{\nu}E(y)\exp\left(\frac{c}{\nu^2}M^2\right),$$

which gives

$$\|Y\|_{L^\infty(0,T;\boldsymbol{H})} + \sqrt{\nu}\|Y\|_{L^2(0,T;\boldsymbol{V})} \le \frac{4\sqrt{2}}{\sqrt{\nu}}\sqrt{E(y)}\exp\left(\frac{c}{\nu^2}M^2\right) \le C(M,\nu)\sqrt{E(y)}.$$

Now

$$\begin{aligned}\|\partial_t Y\|_{L^2(0,T,\boldsymbol{V}')} &\le \big\|\partial_t v + B_1(v) + \nu B_1(Y) + B(Y,y) + B(\bar{y},Y)\big\|_{L^2(0,T;\boldsymbol{V}')} \\ &\le \nu\|Y\|_{L^2(0,T,\boldsymbol{V})} + \|\partial_t v\|_{L^2(0,T,\boldsymbol{V}')} + \|v\|_{L^2(0,T,\boldsymbol{V})} \\ &\quad + c\|Y\|_{L^\infty(0,T;\boldsymbol{H})}\|y\|_{L^2(0,T;\boldsymbol{V})} + c\|\bar{y}\|_{L^\infty(0,T;\boldsymbol{H})}\|Y\|_{L^2(0,T,\boldsymbol{V})} \\ &\le \sqrt{E(y)}\left(2\sqrt{2}\exp\left(\frac{c}{\nu^2}M^2\right) + 2\sqrt{2} + cM\frac{4\sqrt{2}}{\nu}\exp\left(\frac{c}{\nu^2}M^2\right)\right),\end{aligned}$$

and thus

$$\|\partial_t Y\|_{L^2(0,T,\boldsymbol{V}')} \le c(M,\nu)\sqrt{E(y)}. \qquad \square$$

As expected, the constant $c(M,\nu)$ blows up as the viscosity constant $\nu$ goes to zero. The blowup is exponential.

We now proceed with the second part of the proof and would like to show that the only critical points for $E$ correspond to solutions of (11.2). In such a case the search for an element $y$ solution of (11.2) is reduced to the minimization of $E$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

For any $y \in \mathcal{A}$, we now look for a solution $Y_1 \in \mathcal{A}_0$ of the equation

$$\begin{cases} \dfrac{d}{dt}\displaystyle\int_\Omega Y_1 \cdot w + \nu \int_\Omega \nabla Y_1 \cdot \nabla w + \int_\Omega y \cdot \nabla Y_1 \cdot w \\ \quad + \displaystyle\int_\Omega Y_1 \cdot \nabla y \cdot w = -\frac{d}{dt}\int_\Omega v \cdot w - \int_\Omega \nabla v \cdot \nabla w \quad \forall w \in \boldsymbol{V}, \\ Y_1(0) = 0, \end{cases} \tag{11.75}$$

where $v \in \mathcal{A}_0$ is the corrector (associated to $y$) solution of (11.69). The solution $Y_1$ enjoys the following property.

**Proposition 11.3.11.** *For all $y \in \mathcal{A}$, there exists a unique solution $Y_1 \in \mathcal{A}_0$ of* (11.75). *Moreover, if for some $M \in \mathbb{R}$, $\|\partial_t y\|_{L^2(0,T,\boldsymbol{V}')} \leq M$ and $\sqrt{\nu}\|\nabla y\|_{L^2(Q_T)^4} \leq M$, then this solution satisfies*

$$\|\partial_t Y_1\|_{L^2(0,T,\boldsymbol{V}')} + \sqrt{\nu}\|\nabla Y_1\|_{L^2(Q_T)^4} \leq c(M,\nu)\sqrt{E(y)}$$

*for some constant $c(M,\nu) > 0$.*

*Proof.* As in Proposition 11.3.10, (11.75) can be written as

$$\begin{cases} \dfrac{d}{dt}\displaystyle\int_\Omega Y_1 \cdot w + \nu \int_\Omega \nabla Y_1 \cdot \nabla w + \int_\Omega y \cdot \nabla Y_1 \cdot w + \int_\Omega Y_1 \cdot \nabla y \cdot w \\ \quad = -\langle \partial_t v + B_1(v), w\rangle_{\boldsymbol{V}'\times\boldsymbol{V}} \quad \forall w \in \boldsymbol{V}, \\ Y_1(0) = 0. \end{cases} \tag{11.76}$$

Equation (11.76) admits a unique solution $Y_1 \in \mathcal{A}_0$. Indeed, let $y_1 \in L^2(0,T;\boldsymbol{V}) \cap \mathcal{C}([0,T];\boldsymbol{H})$. Moreover, there exists (see [25]) a unique solution $z_1 \in \mathcal{A}_0$ of

$$\begin{cases} \dfrac{d}{dt}\displaystyle\int_\Omega z_1 \cdot w + \nu \int_\Omega \nabla z_1 \cdot \nabla w + \int_\Omega y \cdot \nabla z_1 \cdot w + \int_\Omega y_1 \cdot \nabla y \cdot w \\ \quad = -\langle \partial_t v + B_1(v), w\rangle_{\boldsymbol{V}'\times\boldsymbol{V}} \quad \forall w \in \boldsymbol{V}, \\ z_1(0) = 0. \end{cases} \tag{11.77}$$

Let $\mathcal{T} : y_1 \mapsto z_1$. Then if $z_2 = \mathcal{T}(y_2)$, then $z_1 - z_2$ is a solution of

$$\begin{cases} \dfrac{d}{dt}\displaystyle\int_\Omega (z_1 - z_2) \cdot w + \nu \int_\Omega \nabla(z_1 - z_2) \cdot \nabla w + \int_\Omega y \cdot \nabla(z_1 - z_2) \cdot w \\ \quad + \displaystyle\int_\Omega (y_1 - y_2) \cdot \nabla y \cdot w = 0 \quad \forall w \in \boldsymbol{V}, \\ (z_1 - z_2)(0) = 0, \end{cases}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

and thus, for $w = z_1 - z_2$,

$$\frac{1}{2}\frac{d}{dt}\int_\Omega |z_1 - z_2|^2 + \nu \int_\Omega \left|\nabla(z_1 - z_2)\right|^2 = -\int_\Omega (y_1 - y_2)\cdot\nabla y\cdot(z_1 - z_2).$$

However,

$$\begin{aligned}\left|\int_\Omega (y_1 - y_2)\cdot\nabla y\cdot(z_1 - z_2)\right| &\le c\|y_1 - y_2\|_2\|y\|_{\boldsymbol{V}}\left\|\nabla(z_1 - z_2)\right\|_2 \\ &\le c\|y_1 - y_2\|_2^2\|y\|_{\boldsymbol{V}}^2 + \frac{\nu}{2}\left\|\nabla(z_1 - z_2)\right\|_2^2,\end{aligned}$$

so that

$$\frac{d}{dt}\int_\Omega |z_1 - z_2|^2 + \nu\int_\Omega \left|\nabla(z_1 - z_2)\right|^2 \le c\|y_1 - y_2\|_2^2\|y\|_{\boldsymbol{V}}^2,$$

and for all $t \in [0, T]$,

$$\|z_1 - z_2\|_{L^\infty(0,t,\boldsymbol{H})}^2 + \nu\int_0^t\int_\Omega \left|\nabla(z_1 - z_2)\right|^2 \le c\|y_1 - y_2\|_{L^\infty(0,t,\boldsymbol{H})}^2\int_0^t \|y\|_{\boldsymbol{V}}^2.$$

Since $y \in L^2(0, T; \boldsymbol{V})$, there exists $t' \in ]0, T]$ such that $\int_0^{t'} \|y\|_{\boldsymbol{V}}^2 \le \frac{1}{2c}$. We then have

$$\|z_1 - z_2\|_{L^\infty(0,t',\boldsymbol{H})}^2 + \nu\int_0^{t'}\int_\Omega \left|\nabla(z_1 - z_2)\right|^2 \le \frac{1}{2}\|y_1 - y_2\|_{L^\infty(0,t',\boldsymbol{H})}^2,$$

and the map $\mathcal{T}$ is a contraction mapping on $X = \mathcal{C}([0, t']; \boldsymbol{H}) \cap L^2(0, t'; \boldsymbol{V})$. So $\mathcal{T}$ admits a unique fixed point $Y_1 \in X$. Moreover, from (11.77) we deduce that $\partial_t Y_1 \in L^2(0, t', \boldsymbol{V}')$. Since the map $t \mapsto \int_0^t \|\nabla y\|_2^2$ is a uniformly continuous function, we can take $t' = T$.

For this solution, since $\int_{Q_t} y\cdot\nabla Y_1\cdot Y_1 = 0$, for all $t \in [0, T]$, we have

$$\frac{1}{2}\int_\Omega \left|Y_1(t)\right|^2 + \nu\int_{Q_t} |\nabla Y_1|^2 = -\int_0^t \left\langle B(Y_1, y) + \partial_t v + B_1(v), Y_1\right\rangle_{\boldsymbol{V}'\times\boldsymbol{V}}.$$

Moreover, as in the proof of Proposition 11.3.10, we have

$$\int_\Omega \left|Y_1(t)\right|^2 + \nu\int_{Q_t} |\nabla Y_1|^2 \le \frac{8}{\nu}E(y)\exp\left(\frac{c}{\nu}\int_0^t \|y\|_{\boldsymbol{V}}^2\right), \tag{11.78}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

and thus

$$\begin{aligned}\sqrt{\nu}\|Y_1\|_{L^2(0,T;\boldsymbol{V})} &\leq \frac{2\sqrt{2}}{\sqrt{\nu}}\sqrt{E(y)}\exp\left(\frac{c}{\nu}\int_0^T \|y\|_{\boldsymbol{V}}^2\right)\\ &\leq \frac{2\sqrt{2}}{\sqrt{\nu}}\sqrt{E(y)}\exp\left(\frac{c}{\nu^2}M^2\right) \leq c(M,\nu)\sqrt{E(y)}\end{aligned}$$

and

$$\begin{aligned}&\|\partial_t Y_1\|_{L^2(0,T,\boldsymbol{V}')}\\ &\leq \sqrt{E(y)}\left(2\sqrt{2}\exp\left(\frac{c}{\nu}\int_0^T \|y\|_{\boldsymbol{V}}^2\right) + 2\sqrt{2} + c\|y\|_{L^2(0,T;\boldsymbol{V})}\frac{2\sqrt{2}}{\sqrt{\nu}}\exp\left(\frac{c}{\nu}\int_0^T \|y\|_{\boldsymbol{V}}^2\right)\right.\\ &\quad \left. + c\|y\|_{L^\infty(0,T;\boldsymbol{H})}\frac{2\sqrt{2}}{\nu}\exp\left(\frac{c}{\nu}\int_0^T \|y\|_{\boldsymbol{V}}^2\right)\right)\\ &\leq \sqrt{E(y)}\left(2\sqrt{2}\exp\left(\frac{c}{\nu^2}M^2\right) + 2\sqrt{2} + cM\frac{4\sqrt{2}}{\nu}\exp\left(\frac{c}{\nu^2}M^2\right)\right) \leq c(M,\nu)\sqrt{E(y)}.\end{aligned} \tag{11.79}$$

□

**Proposition 11.3.12.** *For all $y \in \mathcal{A}$, the map $Y \mapsto E(y+Y)$ is a differentiable function on the Hilbert space $\mathcal{A}_0$, and for any $Y \in \mathcal{A}_0$, we have*

$$E'(y)\cdot Y = \langle v, V\rangle_{\mathcal{A}_0} = \int_0^T \langle v, V\rangle_{\boldsymbol{V}} + \int_0^T \langle \partial_t v, \partial_t V\rangle_{\boldsymbol{V}'},$$

*where $V \in \mathcal{A}_0$ is the unique solution in $\mathcal{D}'(0,T)$ of*

$$\begin{cases}\dfrac{d}{dt}\displaystyle\int_\Omega V\cdot w + \int_\Omega \nabla V\cdot\nabla w + \frac{d}{dt}\int_\Omega Y\cdot w + \nu\int_\Omega \nabla Y\cdot\nabla w\\ \qquad + \displaystyle\int_\Omega y\cdot\nabla Y\cdot w + \int_\Omega Y\cdot\nabla y\cdot w = 0 \quad \forall w\in\boldsymbol{V},\\ V(0)=0.\end{cases} \tag{11.80}$$

*Proof.* Let $y \in \mathcal{A}$ and $Y \in \mathcal{A}_0$. We have $E(y+Y) = \frac{1}{2}\|\overline{V}\|_{\mathcal{A}_0}^2$, where $\overline{V} \in \mathcal{A}_0$ is the unique solution of

$$\begin{cases}\dfrac{d}{dt}\displaystyle\int_\Omega \overline{V}\cdot w + \int_\Omega \nabla\overline{V}\cdot\nabla w + \frac{d}{dt}\int_\Omega (y+Y)\cdot w + \nu\int_\Omega \nabla(y+Y)\cdot\nabla w\\ \qquad + \displaystyle\int_\Omega (y+Y)\cdot\nabla(y+Y)\cdot w - \langle f, w\rangle_{\boldsymbol{V}'\times\boldsymbol{V}} = 0 \quad \forall w\in\boldsymbol{V},\\ \overline{V}(0)=0.\end{cases}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

If $v \in \mathcal{A}_0$ is the solution of (11.69) associated with $y$, then $v' \in \mathcal{A}_0$ is the unique solution of

$$\begin{cases} \dfrac{d}{dt}\displaystyle\int_\Omega v' \cdot w + \int_\Omega \nabla v' \cdot \nabla w + \int_\Omega Y \cdot \nabla Y \cdot w = 0 \quad \forall w \in \boldsymbol{V}, \\ v'(0) = 0 \end{cases}$$

and $V \in \mathcal{A}_0$ is the unique solution of (11.80), then it is straightforward to check that $\overline{V} - v - v' - V \in \mathcal{A}_0$ is a solution of

$$\begin{cases} \dfrac{d}{dt}\displaystyle\int_\Omega (\overline{V} - v - v' - V) \cdot w + \int_\Omega \nabla(\overline{V} - v - v' - V) \cdot \nabla w = 0 \quad \forall w \in \boldsymbol{V}, \\ (\overline{V} - v - v' - V)(0) = 0, \end{cases}$$

and therefore $\overline{V} - v - v' - V = 0$. Thus

$$\begin{aligned} E(y+Y) &= \frac{1}{2}\|v + v' + V\|^2_{\mathcal{A}_0} \\ &= \frac{1}{2}\|v\|^2_{\mathcal{A}_0} + \frac{1}{2}\|v'\|^2_{\mathcal{A}_0} + \frac{1}{2}\|V\|^2_{\mathcal{A}_0} + \langle V, v' \rangle_{\mathcal{A}_0} + \langle V, v \rangle_{\mathcal{A}_0} + \langle v, v' \rangle_{\mathcal{A}_0}. \end{aligned}$$

We deduce from (11.80) and (11.71) that

$$\begin{aligned} \|V\|^2_{L^2(0,T,\boldsymbol{V})} \le c\big(&\|\partial_t Y\|^2_{L^2(0,T,\boldsymbol{V}')} + \nu^2 \|B_1(Y)\|^2_{L^2(0,T,\boldsymbol{V}')} \\ &+ \|B(y,Y)\|^2_{L^2(0,T,\boldsymbol{V}')} + \|B(Y,y)\|^2_{L^2(0,T,\boldsymbol{V}')}\big) \end{aligned}$$

and from (11.66), (11.64), and (11.68) that

$$\|V\|^2_{L^2(0,T,\boldsymbol{V})} \le c\|Y\|^2_{\mathcal{A}_0}.$$

Similarly, we deduce from (11.72) that

$$\|\partial_t V\|^2_{L^2(0,T,\boldsymbol{V}')} \le c\|Y\|^2_{\mathcal{A}_0}.$$

Thus

$$\|V\|^2_{\mathcal{A}_0} \le c\|Y\|^2_{\mathcal{A}_0} = o\big(\|Y\|_{\mathcal{A}_0}\big).$$

From (11.71), (11.72), and (11.64) we also deduce that

$$\|v'\|^2_{L^2(0,T,\boldsymbol{V})} \le \|B(Y,Y)\|^2_{L^2(0,T,\boldsymbol{V}')} \le c\|Y\|^2_{L^\infty(0,T,\boldsymbol{H})}\|Y\|^2_{L^2(0,T,\boldsymbol{V})} \le c\|Y\|^4_{\mathcal{A}_0}$$

and

$$\|\partial_t v'\|^2_{L^2(0,T,\boldsymbol{V}')} \le c\|Y\|^2_{L^\infty(0,T,\boldsymbol{H})}\|Y\|^2_{L^2(0,T,\boldsymbol{V})} \le c\|Y\|^4_{\mathcal{A}_0}.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Thus we also have

$$\|v'\|_{\mathcal{A}_0}^2 \leq c\|Y\|_{\mathcal{A}_0}^4 = o(\|Y\|_{\mathcal{A}_0}).$$

From the previous estimates we obtain

$$\left|\langle V, v'\rangle_{\mathcal{A}_0}\right| \leq \|V\|_{\mathcal{A}_0}\|v'\|_{\mathcal{A}_0} \leq c\|Y\|_{\mathcal{A}_0}^3 = o(\|Y\|_{\mathcal{A}_0})$$

and

$$\left|\langle v, v'\rangle_{\mathcal{A}_0}\right| \leq \|v\|_{\mathcal{A}_0}\|v'\|_{\mathcal{A}_0} \leq c\sqrt{E(y)}\|Y\|_{\mathcal{A}_0}^2 = o(\|Y\|_{\mathcal{A}_0}),$$

and thus

$$E(y+Y) = E(y) + \langle v, V\rangle_{\mathcal{A}_0} + o(\|Y\|_{\mathcal{A}_0}).$$

Eventually, the estimate

$$\left|\langle v, V\rangle_{\mathcal{A}_0}\right| \leq \|v\|_{\mathcal{A}_0}\|V\|_{\mathcal{A}_0} \leq c\sqrt{E(y)}\|Y\|_{\mathcal{A}_0}$$

gives the continuity of the linear map $Y \mapsto \langle v, V\rangle_{\mathcal{A}_0}$. □

We are now in position to prove the following result.

**Proposition 11.3.13.** *If $(y_k)_{k\in\mathbb{N}}$ is a sequence of $\mathcal{A}$ bounded in $L^2(0,T;\boldsymbol{V}) \cap H^1(0,T;\boldsymbol{V}')$ satisfying $E'(y_k) \to 0$ as $k \to \infty$, then $E(y_k) \to 0$ as $k \to \infty$.*

*Proof.* For any $y \in \mathcal{A}$ and $Y \in \mathcal{A}_0$, we have

$$E'(y)\cdot Y = \langle v, V\rangle_{\mathcal{A}_0} = \int_0^T \langle v, V\rangle_{\boldsymbol{V}} + \int_0^T \langle \partial_t v, \partial_t V\rangle_{\boldsymbol{V}'},$$

where $V \in \mathcal{A}_0$ is the unique solution in $\mathcal{D}'(0,T)$ of (11.80). In particular, taking $Y = Y_1$ defined by (11.75), we define a solution $V_1$ of

$$\begin{cases} \dfrac{d}{dt}\displaystyle\int_\Omega V_1\cdot w + \int_\Omega \nabla V_1\cdot\nabla w + \frac{d}{dt}\int_\Omega Y_1\cdot w + \nu\int_\Omega \nabla Y_1\cdot\nabla w + \int_\Omega y\cdot\nabla Y_1\cdot w \\ \qquad + \displaystyle\int_\Omega Y_1\cdot\nabla y\cdot w = 0 \quad \forall w \in \boldsymbol{V}, \\ V_1(0) = 0. \end{cases} \tag{11.81}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Summing (11.81) and (11.75), we obtain that $V_1 - v$ solves (11.70) with $F \equiv 0$ and $z_0 = 0$. This implies that $V_1$ and $v$ coincide, and then

$$E'(y) \cdot Y_1 = \int_0^T \|v\|_{\boldsymbol{V}}^2 + \int_0^T \|\partial_t v\|_{\boldsymbol{V}'}^2 = 2E(y) \quad \forall y \in \mathcal{A}. \tag{11.82}$$

Let now, for any $k \in \mathbb{N}$, $Y_{1,k}$ be the solution of (11.75) associated with $y_k$. The previous equality writes $E'(y_k) \cdot Y_{1,k} = 2E(y_k)$ and implies our statement, since by Proposition 11.3.11, $Y_{1,k}$ is uniformly bounded in $\mathcal{A}_0$. □

### 11.3.3 Minimizing sequence for $E$

As in the previous section, equality (11.82) shows that $-Y_1$ given by the solution of (11.75) is a descent direction for the functional $E$. Remark also that, in view of (11.75), the corrector $V$ associated with $Y_1$, given by (11.80) with $Y = Y_1$, is nothing else than the corrector $v$ itself. Therefore we can define, for any $m \geq 1$, a minimizing sequence $\{y_k\}_{(k\in\mathbb{N})}$ for $E$ as follows:

$$\begin{cases} y_0 \in \mathcal{A}, \\ y_{k+1} = y_k - \lambda_k Y_{1,k}, \quad k \geq 0, \\ E(y_k - \lambda_k Y_{1,k}) = \min\limits_{\lambda\in[0,m]} E(y_k - \lambda Y_{1,k}) \end{cases} \tag{11.83}$$

with the solution $Y_{1,k} \in \mathcal{A}_0$ of the equation

$$\begin{cases} \dfrac{d}{dt}\displaystyle\int_\Omega Y_{1,k} \cdot w + \nu \int_\Omega \nabla Y_{1,k} \cdot \nabla w + \int_\Omega y_k \cdot \nabla Y_{1,k} \cdot w \\ \qquad + \displaystyle\int_\Omega Y_{1,k} \cdot \nabla y_k \cdot w = -\frac{d}{dt}\int_\Omega v_k \cdot w - \int_\Omega \nabla v_k \cdot \nabla w \quad \forall w \in \boldsymbol{V}. \\ Y_{1,k}(0) = 0, \end{cases} \tag{11.84}$$

where $v_k \in \mathcal{A}_0$ is the corrector solution (associated with $y_k$) of (11.69), leading (see (11.82)) to $E'(y_k) \cdot Y_{1,k} = 2E(y_k)$. For any $k > 0$, the direction $Y_{1,k}$ vanishes when $E(y_k)$ vanishes.

**Lemma 11.3.14.** *Let $(y_k)_{k\in\mathbb{N}}$ the sequence of $\mathcal{A}$ defined by* (11.83). *Then $(y_k)_{k\in\mathbb{N}}$ is a bounded sequence of $H^1(0,T;\boldsymbol{V}') \cap L^2(0,T;\boldsymbol{V})$, and $(E(y_k))_{k\in\mathbb{N}}$ is a decreasing sequence.*

*Proof.* From (11.83) we deduce that, for all $k \in \mathbb{N}$,

$$E(y_{k+1}) = E(y_k - \lambda_k Y_{1,k}) = \min_{\lambda\in[0,m]} E(y_k - \lambda Y_{1,k}) \leq E(y_k),$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

and thus the sequence $(E(y_k))_{k\in\mathbb{N}}$ decreases, and $E(y_k) \le E(y_0)$ for all $k \in \mathbb{N}$. Moreover, from the construction of the corrector $v_k \in \mathcal{A}_0$ associated with $y_k \in \mathcal{A}$ given by (11.69) we deduce from Proposition 11.3.5 that $y_k \in \mathcal{A}$ is the unique solution of

$$\begin{cases} \dfrac{d}{dt}\displaystyle\int_\Omega y_k \cdot w + \nu \int_\Omega \nabla y_k \cdot \nabla w + \int_\Omega y_k \cdot \nabla y_k \cdot w \\ \qquad = \langle f, w\rangle_{\boldsymbol{V}'\times\boldsymbol{V}} - \dfrac{d}{dt}\displaystyle\int_\Omega v_k \cdot w - \int_\Omega \nabla v_k \cdot \nabla w \quad \forall w \in \boldsymbol{V}, \\ y_k(0) = u_0, \end{cases}$$

and using (11.64) and (11.66), we get

$$\begin{aligned} \|y_k\|^2_{L^\infty(0,T;\boldsymbol{H})} &\le \|u_0\|^2_{\boldsymbol{H}} + \frac{1}{\nu}\left\|f - \partial_t v_k - B_1(v_k)\right\|^2_{L^2(0,T;\boldsymbol{V}')} \\ &\le \|u_0\|^2_{\boldsymbol{H}} + \frac{2}{\nu}\|f\|^2_{L^2(0,T;\boldsymbol{V}')} + \frac{2}{\nu}\|\partial_t v_k\|^2_{L^2(0,T;\boldsymbol{V}')} + \frac{2}{\nu}\|v_k\|^2_{L^2(0,T;\boldsymbol{V})} \\ &\le \|u_0\|^2_{\boldsymbol{H}} + \frac{2}{\nu}\|f\|^2_{L^2(0,T;\boldsymbol{V}')} + \frac{4}{\nu}E(y_k) \\ &\le \|u_0\|^2_{\boldsymbol{H}} + \frac{2}{\nu}\|f\|^2_{L^2(0,T;\boldsymbol{V}')} + \frac{4}{\nu}E(y_0), \end{aligned} \tag{11.85}$$

$$\begin{aligned} \nu\|y_k\|^2_{L^2(0,T;\boldsymbol{V})} &\le \|u_0\|^2_{\boldsymbol{H}} + \frac{1}{\nu}\left\|f - \partial_t v_k - B_1(v_k)\right\|^2_{L^2(0,T;\boldsymbol{V}')} \\ &\le \|u_0\|^2_{\boldsymbol{H}} + \frac{2}{\nu}\|f\|^2_{L^2(0,T;\boldsymbol{V}')} + \frac{4}{\nu}E(y_0), \end{aligned} \tag{11.86}$$

and

$$\begin{aligned} \|\partial_t y_k\|_{L^2(0,T;\boldsymbol{V}')} &\le \left\|f - \partial_t v_k - B_1(v_k) - B(y_k, y_k) - \nu B_1(y_k)\right\|_{L^2(0,T;\boldsymbol{V}')} \\ &\le \|f\|_{L^2(0,T;\boldsymbol{V}')} + \|\partial_t v_k\|_{L^2(0,T;\boldsymbol{V}')} + \|v_k\|_{L^2(0,T;\boldsymbol{V})} \\ &\quad + c\|y_k\|_{L^\infty(0,T;\boldsymbol{H})}\|y_k\|_{L^2(0,T;\boldsymbol{V})} + \nu\|y_k\|_{L^2(0,T;\boldsymbol{V})} \\ &\le \|f\|_{L^2(0,T;\boldsymbol{V}')} + 2\sqrt{E(y_k)} + \sqrt{\nu}\|u_0\|_{\boldsymbol{H}} + \sqrt{2}\|f\|_{L^2(0,T;\boldsymbol{V}')} \\ &\quad + 2\sqrt{E(y_0)} + c\left(\|u_0\|^2_{\boldsymbol{H}} + \frac{2}{\nu}\|f\|^2_{L^2(0,T;\boldsymbol{V}')} + \frac{4}{\nu}E(y_0)\right) \\ &\le 3\|f\|_{L^2(0,T;\boldsymbol{V}')} + 4\sqrt{E(y_0)} + \sqrt{\nu}\|u_0\|_{\boldsymbol{H}} \\ &\quad + c\left(\|u_0\|^2_{\boldsymbol{H}} + \frac{2}{\nu}\|f\|^2_{L^2(0,T;\boldsymbol{V}')} + \frac{4}{\nu}E(y_0)\right). \end{aligned}$$

□

**Lemma 11.3.15.** *Let $\{y_k\}_{k\in\mathbb{N}}$ be the sequence of $\mathcal{A}$ defined by* (11.83)*. Then for all $\lambda \in [0, m]$, we have the following estimate:*

$$E(y_k - \lambda Y_{1,k}) \le E(y_k)\left(|1-\lambda| + \lambda^2 \frac{c}{\nu\sqrt{\nu}}\sqrt{E(y_k)}\exp\left(\frac{c}{\nu}\int_0^T \|y_k\|^2_{\boldsymbol{V}}\right)\right)^2. \tag{11.87}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*Proof.* Let $V_k$ be the corrector associated with $y_k - \lambda Y_{1,k}$. It is easy to check that $V_k$ is given by $(1-\lambda)v_k + \lambda^2 \bar{\bar{v}}_k$, where $\bar{\bar{v}}_k \in \mathcal{A}_0$ solves

$$\begin{cases} \dfrac{d}{dt}\displaystyle\int_\Omega \bar{\bar{v}}_k \cdot w + \int_\Omega \nabla \bar{\bar{v}}_k \cdot \nabla w + \int_\Omega Y_{1,k} \cdot \nabla Y_{1,k} \cdot w = 0 \quad \forall w \in \boldsymbol{V}, \\ \bar{\bar{v}}_k(0) = 0, \end{cases} \tag{11.88}$$

and thus

$$\begin{aligned} 2E(y_k - \lambda Y_{1,k}) = \|V_k\|^2_{\mathcal{A}_0} &= \|(1-\lambda)v_k + \lambda^2 \bar{\bar{v}}_k\|^2_{\mathcal{A}_0} \\ &\le \left(|1-\lambda| \|v_k\|_{\mathcal{A}_0} + \lambda^2 \|\bar{\bar{v}}_k\|_{\mathcal{A}_0}\right)^2 \\ &\le \left(\sqrt{2}|1-\lambda| \sqrt{E(y_k)} + \lambda^2 \|\bar{\bar{v}}_k\|_{\mathcal{A}_0}\right)^2, \end{aligned} \tag{11.89}$$

which gives

$$E(y_k - \lambda Y_{1,k}) \le \left(|1-\lambda| \sqrt{E(y_k)} + \frac{\lambda^2}{\sqrt{2}} \|\bar{\bar{v}}_k\|_{\mathcal{A}_0}\right)^2 := g(\lambda, y_k). \tag{11.90}$$

From (11.88), (11.71), (11.72), and (11.64) we deduce that

$$\|\bar{\bar{v}}_k\|_{L^2(0,T;\boldsymbol{V})} \le \left\|B(Y_{1,k}, Y_{1,k})\right\|_{L^2(0,T;\boldsymbol{V}')} \le c \|Y_{1,k}\|_{L^\infty(0,T;\boldsymbol{H})} \|Y_{1,k}\|_{L^2(0,T;\boldsymbol{V})}$$

and

$$\begin{aligned} \|\partial_t \bar{\bar{v}}_k\|_{L^2(0,T;\boldsymbol{V}')} &\le \left\|-B_1(\bar{\bar{v}}_k) - B(Y_{1,k}, Y_{1,k})\right\|_{L^2(0,T;\boldsymbol{V}')} \\ &\le \|\bar{\bar{v}}_k\|_{L^2(0,T;\boldsymbol{V})} + c \|Y_{1,k}\|_{L^\infty(0,T;\boldsymbol{H})} \|Y_{1,k}\|_{L^2(0,T;\boldsymbol{V})} \\ &\le c \|Y_{1,k}\|_{L^\infty(0,T;\boldsymbol{H})} \|Y_{1,k}\|_{L^2(0,T;\boldsymbol{V})}. \end{aligned}$$

On the other hand, from (11.78) we deduce that

$$\|Y_{1,k}\|^2_{L^\infty(0,T;\boldsymbol{H})} + \nu \|Y_{1,k}\|^2_{L^2(0,T;\boldsymbol{V})} \le \frac{16}{\nu} E(y_k) \exp\left(\frac{c}{\nu} \int_0^T \|y_k\|^2_{\boldsymbol{V}}\right). \tag{11.91}$$

Thus

$$\|\bar{\bar{v}}_k\|_{L^2(0,T;\boldsymbol{V})} \le \frac{c}{\nu\sqrt{\nu}} E(y_k) \exp\left(\frac{c}{\nu} \int_0^T \|y_k\|^2_{\boldsymbol{V}}\right)$$

and

$$\|\partial_t \bar{\bar{v}}_k\|_{L^2(0,T;\boldsymbol{V}')} \le \frac{c}{\nu\sqrt{\nu}} E(y_k) \exp\left(\frac{c}{\nu} \int_0^T \|y_k\|^2_{\boldsymbol{V}}\right),$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

which gives

$$\|\overline{\overline{v}}_k\|_{\mathcal{A}_0} = \sqrt{\|\overline{\overline{v}}_k\|^2_{L^2(0,T;\boldsymbol{V})} + \|\partial_t \overline{\overline{v}}_k\|^2_{L^2(0,T;\boldsymbol{V})}} \leq \frac{c}{\nu\sqrt{\nu}} E(y_k) \exp\left(\frac{c}{\nu}\int_0^T \|y_k\|^2_{\boldsymbol{V}}\right).$$

Then from (11.90) we deduce (11.87). □

**Lemma 11.3.16.** *Let $(y_k)_{k\in\mathbb{N}}$ be the sequence of $\mathcal{A}$ defined by* (11.83). *Then $E(y_k) \to 0$ as $k \to \infty$. Moreover, there exists $k_0 \in \mathbb{N}$ such that the sequence $(E(y_k))_{k\geq k_0}$ decays quadratically.*

*Proof.* Using (11.86), we deduce from (11.87) that for all $\lambda \in [0, m]$ and $k \in \mathbb{N}^\star$,

$$\sqrt{E(y_{k+1})} \leq \sqrt{E(y_k)}\left(|1-\lambda| + \lambda^2 C_1 \sqrt{E(y_k)}\right), \tag{11.92}$$

where $C_1 = \frac{c}{\nu\sqrt{\nu}} \exp(\frac{c}{\nu^2}\|u_0\|^2_{\boldsymbol{H}} + \frac{c}{\nu^3}\|f\|^2_{L^2(0,T;\boldsymbol{V}')} + \frac{c}{\nu^3}E(y_0))$ does not depend on $y_k$.

Let us define the function $p_k(\lambda) = |1-\lambda| + \lambda^2 C_1\sqrt{E(y_k)}$ for $\lambda \in [0, m]$. If $C_1\sqrt{E(y_0)} < 1$ (and thus $C_1\sqrt{E(y_k)} < 1$ for all $k \in \mathbb{N}$), then

$$\min_{\lambda\in[0,m]} p_k(\lambda) \leq p_k(1) = C_1\sqrt{E(y_k)},$$

and thus

$$C_1\sqrt{E(y_{k+1})} \leq \left(C_1\sqrt{E(y_k)}\right)^2, \tag{11.93}$$

implying that $C_1\sqrt{E(y_k)} \to 0$ as $k \to \infty$ with quadratic rate.

Suppose now that $C_1\sqrt{E(y_0)} \geq 1$ and denote $I = \{k \in \mathbb{N},\ C_1\sqrt{E(y_k)} \geq 1\}$. Let us prove that $I$ is a finite subset of $\mathbb{N}$. For all $k \in I$, since $C_1\sqrt{E(y_k)} \geq 1$,

$$\begin{aligned}\min_{\lambda\in[0,m]} p_k(\lambda) &= \min_{\lambda\in[0,1]} p_k(\lambda) = p_k\left(\frac{1}{2C_1\sqrt{E(y_k)}}\right)\\ &= 1 - \frac{1}{4C_1\sqrt{E(y_k)}} \leq 1 - \frac{1}{4C_1\sqrt{E(y_0)}} < 1,\end{aligned}$$

and thus, for all $k \in I$,

$$\sqrt{E(y_{k+1})} \leq \left(1 - \frac{1}{4C_1\sqrt{E(y_0)}}\right)\sqrt{E(y_k)} \leq \left(1 - \frac{1}{4C_1\sqrt{E(y_0)}}\right)^{k+1}\sqrt{E(y_0)}.$$

Since $(1 - \frac{1}{4C_1\sqrt{E(y_0)}})^{k+1} \to 0$ as $k \to +\infty$, there exists $k_0 \in \mathbb{N}$ such that $C_1\sqrt{E(y_{k+1})} < 1$ for all $k \geq k_0$. Thus $I$ is a finite subset of $\mathbb{N}$. Arguing as in the first case, it follows that $C_1\sqrt{E(y_k)} \to 0$ as $k \to \infty$. □

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Remark 11.3.17.** In view of the estimate above of the constant $C_1$, the number of iterates $k_1$ necessary to achieve the quadratic regime (from which the convergence is very fast) is of the order $\nu^{-3/2}\exp(\frac{c}{\nu^2}\|u_0\|^2_{\boldsymbol{H}} + \frac{c}{\nu^3}\|f\|^2_{L^2(0,T;\boldsymbol{V}')} + \frac{c}{\nu^3}E(y_0))$ and therefore increases rapidly as $\nu \to 0$. To reduce the effect of the term $\nu^{-3}E(y_0)$, it is thus relevant, for small values of $\nu$, to couple the algorithm with a continuation approach with respect to $\nu$ (i. e., start the sequence $(y_k)_{(k\geq 0)}$ with a solution $y_0$ of (11.2) associated with $\bar{\nu} > \nu$ so that $\nu^{-3}E(y_0)$ is at most of order $\mathcal{O}(\nu^{-2})$).

**Lemma 11.3.18.** *Let $(y_k)_{k\in\mathbb{N}}$ the sequence of $\mathcal{A}$ defined by* (11.83). *Then $\lambda_k \to 1$ as $k \to \infty$.*

*Proof.* We have

$$\begin{aligned} 2E(y_{k+1}) &= 2E(y_k - \lambda_k Y_{1,k}) \\ &= (1-\lambda_k)^2\|v_k\|^2_{\mathcal{A}_0} + 2\lambda_k^2(1-\lambda_k)\langle v_k, \bar{\bar{v}}_k\rangle_{\mathcal{A}_0} + \lambda_k^4\|\bar{\bar{v}}_k\|^2_{\mathcal{A}_0} \\ &= 2(1-\lambda_k)^2E(y_k) + 2\lambda_k^2(1-\lambda_k)\langle v_k, \bar{\bar{v}}_k\rangle_{\mathcal{A}_0} + \lambda_k^4\|\bar{\bar{v}}_k\|^2_{\mathcal{A}_0}, \end{aligned}$$

and thus, as long as $E(y_k) \neq 0$,

$$(1-\lambda_k)^2 = \frac{E(y_{k+1})}{E(y_k)} - \lambda_k^2(1-\lambda_k)\frac{\langle v_k, \bar{\bar{v}}_k\rangle_{\mathcal{A}_0}}{E(y_k)} - \lambda_k^4\frac{\|\bar{\bar{v}}_k\|^2_{\mathcal{A}_0}}{2E(y_k)}.$$

Since

$$\|\bar{\bar{v}}_k\|_{\mathcal{A}_0} \leq \frac{c}{\nu\sqrt{\nu}}E(y_k)\exp\left(\frac{c}{\nu}\int_0^T \|y_k\|^2_{\boldsymbol{V}}\right),$$

we then have

$$\left|\frac{\langle v_k, \bar{\bar{v}}_k\rangle_{\mathcal{A}_0}}{E(y_k)}\right| \leq \frac{\|v_k\|_{\mathcal{A}_0}\|\bar{\bar{v}}_k\|_{\mathcal{A}_0}}{E(y_k)} \leq \frac{c}{\nu\sqrt{\nu}}\sqrt{E(y_k)}\exp\left(\frac{c}{\nu}\int_0^T \|y_k\|^2_{\boldsymbol{V}}\right) \to 0$$

and

$$0 \leq \frac{\|\bar{\bar{v}}_k\|^2_{\mathcal{A}_0}}{E(y_k)} \leq \frac{c}{\nu^3}E(y_k)\exp\left(\frac{c}{\nu}\int_0^T \|y_k\|^2_{\boldsymbol{V}}\right) \to 0$$

as $k \to +\infty$. Consequently, since $\lambda_k \in [0, m]$ and $\frac{E(y_{k+1})}{E(y_k)} \to 0$, we deduce that $(1-\lambda_k)^2 \to 0$, that is, $\lambda_k \to 1$ as $k \to \infty$. □

From Lemmas 11.3.14 and 11.3.16 and Proposition 11.3.10 we can deduce the following:

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Proposition 11.3.19.** *Let $(y_k)_{k\in\mathbb{N}}$ be the sequence of $\mathcal{A}$ defined by* (11.83)*. Then $y_k \to \bar{y}$ in $H^1(0,T;\boldsymbol{V}') \cap L^2(0,T;\boldsymbol{V})$, where $\bar{y} \in \mathcal{A}$ is the unique solution of* (11.2) *given in Proposition* 11.3.5.

## 11.3.4 Remarks

**Remark 11.3.20.** The strong convergence of the sequence $(y_k)_{k>0}$ is a consequence of the coercivity inequality (11.74), which is itself a consequence of the uniqueness of the solution $y$ of (11.2). In fact, we can directly prove that $(y_k)_{k>0}$ is a convergent sequence in the Hilbert space $\mathcal{A}$ as follows; from (11.83) and the previous proposition we deduce that the series $\sum_{k\geq 0}\lambda_k Y_{1k}$ converges in $H^1(0,T;\boldsymbol{V}')\cap L^2(0,T;\boldsymbol{V})$ and $\bar{y} = y_0 + \sum_{k=0}^{+\infty}\lambda_k Y_{1k}$. Moreover, $\sum \lambda_k \|Y_{1k}\|_{L^2(0,T;\boldsymbol{V})}$ converges, and if we denote by $k_0$ one $k \in \mathbb{N}$ such that $C_1\sqrt{E(y_k)} < 1$ (see Lemma 11.3.16), then for all $k \geq k_0$, using (11.91), (11.86), and (11.93) (since we can choose $C_1 > 1$), we have

$$\begin{aligned}
\|\bar{y} - y_k\|_{L^2(0,T;\boldsymbol{V})} &= \left\| \sum_{i=k+1}^{+\infty} \lambda_i Y_{1i} \right\|_{L^2(0,T;\boldsymbol{V})} \leq \sum_{i=k+1}^{+\infty} \lambda_i \|Y_{1i}\|_{L^2(0,T;\boldsymbol{V})} \\
&\leq m \sum_{i=k+1}^{+\infty} \sqrt{C_1 E(y_i)} \leq m \sum_{i=k+1}^{+\infty} C_1 \sqrt{E(y_i)} \\
&\leq m \sum_{i=k+1}^{+\infty} \left(C_1\sqrt{E(y_{k_0})}\right)^{2^{i-k_0}} \leq m \sum_{i=0}^{+\infty} \left(C_1\sqrt{E(y_{k_0})}\right)^{2^{i+k+1-k_0}} \\
&\leq m\left(C_1\sqrt{E(y_{k_0})}\right)^{2^{k+1-k_0}} \sum_{i=0}^{+\infty} \left(C_1\sqrt{E(y_{k_0})}\right)^{2^i} \\
&\leq cm\left(C_1\sqrt{E(y_{k_0})}\right)^{2^{k+1-k_0}}.
\end{aligned} \tag{11.94}$$

From (11.79) we deduce that

$$\begin{aligned}
&\|\partial_t Y_{1,k}\|_{L^2(0,T,\boldsymbol{V}')} \\
&\leq \sqrt{E(y_k)}\left( 2\sqrt{2}\exp\left(\frac{c}{\nu}\int_0^T \|y_k\|_{\boldsymbol{V}}^2\right) + 2\sqrt{2} + c\|y_k\|_{L^2(0,T;\boldsymbol{V})}\frac{2\sqrt{2}}{\sqrt{\nu}}\exp\left(\frac{c}{\nu}\int_0^T \|y_k\|_{\boldsymbol{V}}^2\right)\right. \\
&\quad \left. + c\|y_k\|_{L^\infty(0,T;\boldsymbol{H})}\frac{2\sqrt{2}}{\nu}\exp\left(\frac{c}{\nu}\int_0^T \|y_k\|_{\boldsymbol{V}}^2\right)\right),
\end{aligned}$$

which gives, using (11.85) and (11.86),

$$\begin{aligned}
&\|\partial_t Y_{1,k}\|_{L^2(0,T,\boldsymbol{V}')} \\
&\leq \sqrt{E(y_k)}\left(2\sqrt{2}\exp\left(\frac{c}{\nu}\int_0^T \|y_k\|_{\boldsymbol{V}}^2\right) + 2\sqrt{2}\right)
\end{aligned} \tag{11.95}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$$+\frac{c}{\nu\sqrt{\nu}}\big(\sqrt{\nu}\|u_0\|_{\boldsymbol{H}}+\sqrt{2}\|f\|_{L^2(0,T;\boldsymbol{V}')}+2\sqrt{E(y_0)}\big)\exp\left(\frac{c}{\nu}\int_0^T\|y_k\|_{\boldsymbol{V}}^2\right)$$
$$\leq C_2\sqrt{E(y_k)},$$

where

$$C_2 = c\exp\left(\frac{c}{\nu^2}\|u_0\|_{\boldsymbol{H}}^2+\frac{c}{\nu^3}\|f\|_{L^2(0,T;\boldsymbol{V}')}^2+\frac{c}{\nu^3}E(y_0)\right)$$
$$\times\left(1+\frac{1}{\nu\sqrt{\nu}}\big(\sqrt{\nu}\|u_0\|_{\boldsymbol{H}}+\sqrt{2}\|f\|_{L^2(0,T;\boldsymbol{V}')}+2\sqrt{E(y_0)}\big)\right).$$

Arguing as in the proof of Lemma 11.3.16, there exists $k_1 \in \mathbb{N}$ such that for all $k \geq k_1$,

$$C_2\sqrt{E(y_{k+1})} \leq \big(C_2\sqrt{E(y_k)}\big)^2,$$

and thus

$$\begin{aligned}\|\partial_t\bar{y}-\partial_t y_k\|_{L^2(0,T;\boldsymbol{V}')} &= \left\|\sum_{i=k+1}^{+\infty}\lambda_i\partial_t Y_{1i}\right\|_{L^2(0,T;\boldsymbol{V}')} \leq \sum_{i=k+1}^{+\infty}\lambda_i\|\partial_t Y_{1i}\|_{L^2(0,T;\boldsymbol{V}')}\\ &\leq m\sum_{i=k+1}^{+\infty}\sqrt{C_2E(y_i)}\\ &\leq m\big(C_2\sqrt{E(y_{k_1})}\big)^{2^{k+1-k_1}}\sum_{i=0}^{+\infty}\big(C_2\sqrt{E(y_{k_1})}\big)^{2^i}\\ &\leq mc\big(C_2\sqrt{E(y_{k_1})}\big)^{2^{k+1-k_1}}.\end{aligned} \tag{11.96}$$

**Remark 11.3.21.** Lemmas 11.3.14, 11.3.15, and 11.3.16 and Proposition 11.3.19 remain true if we replace the minimization of $\lambda$ over $[0,m]$ by the minimization over $\mathbb{R}_+$. However, the sequence $(\lambda_k)_{k>0}$ may not be bounded in $\mathbb{R}_+$ (the fourth-order polynomial $\lambda \to E(y_k-\lambda Y_k)$ may admit a critical point far from 1. In that case, (11.94) and (11.96) may not hold for some $m>0$.

Similarly, Lemmas 11.3.16 and 11.3.18 and Proposition 11.3.19 remain true for the sequence $\{y_k\}_{k\geq 0}$ defined as follows:

$$\begin{cases} y_0 \in \mathcal{A},\\ y_{k+1} = y_k-\lambda_k Y_{1,k}, \quad k\geq 0,\\ g(\lambda_k,y_k) = \min\limits_{\lambda\in\mathbb{R}^+} g(\lambda,y_k),\end{cases} \tag{11.97}$$

leading to $\lambda_k \in ]0,1]$ for all $k\geq 0$ and $\lim_{k\to\infty}\lambda_k \to 1^-$. The fourth-order polynomial $g$ is defined in (11.90).

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Remark 11.3.22.** Let us consider the application $F : \mathcal{A} \to L^2(0,T;\boldsymbol{V}')$ defined as $F(y) = y_t + \nu B_1(y) + B(y,y) - f$. The sequence $(y_k)_{k>0}$ associated with the Newton method to find the zero of $F$ is defined as follows:

$$\begin{cases} y_0 \in \mathcal{A}, \\ F'(y_k)\cdot(y_{k+1} - y_k) = -F(y_k), \quad k \geq 0. \end{cases}$$

As in the previous section, this sequence coincides with the sequence obtained from (11.83) if $\lambda_k$ is fixed equal to one. Algorithm (11.83), which optimizes the parameter $\lambda_k \in [0,m]$, $m \geq 1$, to minimize $E(y_k)$ or, equivalently, $\|F(y_k)\|_{L^2(0,T;\boldsymbol{V}')}$, corresponds therefore to the so-called damped Newton method for the application $F$ (see [4]). As the iterates increase, the optimal parameter $\lambda_k$ converges to one (according to Lemma 11.2.12), and this globally convergent method behaves like the standard Newton method (for which $\lambda_k$ is fixed equal to one). This explains the quadratic rate after a finite number of iterates.

**Remark 11.3.23.** In a different functional framework a similar approach is considered in [18]; more precisely, the author introduces the functional $E : \mathcal{V} \to \mathbb{R}$ defined $E(y) = \frac{1}{2}\|\nabla v\|^2_{L^2(Q_T)}$ with $\mathcal{V} := y_0 + \mathcal{V}_0$, $y_0 \in H^1(Q_T)$, and $\mathcal{V}_0 := \{u \in H^1(Q_T;\mathbb{R}^d), u(t,\cdot) \in \boldsymbol{V}\ \forall t \in (0,T), u(0,\cdot) = 0\}$ where $v(t,\cdot)$ solves, for all $t \in (0,T)$, the steady Navier–Stokes equation with source term $y_t(t,\cdot) - \nu\Delta y(t,\cdot) + (y(t,\cdot)\cdot\nabla)y(t,\cdot) - f(t,\cdot)$. Strong solutions are therefore considered assuming that $u_0 \in \boldsymbol{V}$ and $f \in (L^2(Q_T))^2$. A bound of $E(y)$ implies a bound of $y$ in $L^2(0,T;\boldsymbol{V})$ but not in $H^1(0,T;L^2(\Omega)^2)$. This prevents getting the convergence of minimizing sequences in $\mathcal{V}$.

**Remark 11.3.24.** This approach, which minimizes an appropriate norm of the solution, is refereed to in the literature as a variational approach. We mention notably the work [18], where strong solutions of (11.1) are characterized in the two-dimensional case in terms of the critical points of a quadratic functional close to $\widetilde{E}$. Similarly, the authors in [17] show that the functional

$$I^\epsilon(y) = \int_0^\infty\int_\Omega e^{-t/\epsilon}\left\{|\partial_t y + y\cdot\nabla y|^2 + |y\cdot\nabla y|^2 + \frac{\nu}{\epsilon}|\nabla y|^2\right\}$$

admits minimizers $u^\epsilon$ for all $\epsilon > 0$, and, up to subsequences, such minimizers converge weakly to a Leray–Hopf solution of (11.1) as $\epsilon \to 0$.

# 11.4 Numerical illustrations

## 11.4.1 Algorithm: approximation

We detail the main steps of the iterative algorithm (11.83). First, we define the initial term $y_0$ of the sequence $(y_k)_{(k\geq 0)}$ as the solution of the Stokes problem, solved by the

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

backward Euler scheme

$$\begin{cases} \displaystyle\int_\Omega \frac{y_0^{n+1}-y_0^n}{\delta t}\cdot w + \overline{\nu}\int_\Omega \nabla y_0^{n+1}\cdot\nabla w = \langle f^n, w\rangle_{\boldsymbol{V}'\times\boldsymbol{V}} \quad \forall w\in\boldsymbol{V},\ \forall n\geq 0, \\ y_0^0(\cdot,0) = u_0 \quad \text{in } \Omega \end{cases} \tag{11.98}$$

for some $\overline{\nu} > 0$. The incompressibility constraint is taken into account through a Lagrange multiplier $\lambda \in L^2(\Omega)$, leading to the mixed problem

$$\begin{cases} \displaystyle\int_\Omega \frac{y_0^{n+1}-y_0^n}{\delta t}\cdot w + \nu\int_\Omega \nabla y_0^{n+1}\cdot\nabla w + \int_\Omega \lambda^{n+1}\,\nabla\cdot w = \langle f^n, w\rangle_{\boldsymbol{V}'\times\boldsymbol{V}} \\ \qquad \forall w\in (H_0^1(Q_T))^2,\ \forall n\geq 0, \\ \displaystyle\int_\Omega \mu\,\nabla\cdot y_0^{n+1} = 0, \quad \forall \mu\in L^2(\Omega)\ \forall n\geq 0, \\ y_0^0 = u_0 \quad \text{in } \Omega. \end{cases} \tag{11.99}$$

A conformal approximation in space is used for $(H_0^1(\Omega))^2 \times L^2(\Omega)$ based on the inf–sup stable $\mathbb{P}_2/\mathbb{P}_1$ Taylor–Hood finite element. Then, assuming that (an approximation $(y_{h,k}^n)_{\{n,h\}}$ of) $y_k$ has been obtained for some $k \geq 0$, $y_{k+1}$ is obtained as follows.

(i) From $y_k$, computation of (an approximation of) the corrector $v_k$ through the backward Euler scheme

$$\begin{cases} \displaystyle\int_\Omega \frac{v_k^{n+1}-v_k^n}{\delta t}\cdot w + \int_\Omega \nabla v_k^{n+1}\cdot\nabla w + \int_\Omega \frac{y_k^{n+1}-y_k^n}{\delta t}\cdot w + \nu\int_\Omega \nabla y_k^{n+1}\cdot\nabla w \\ \qquad + \displaystyle\int_\Omega y_k^{n+1}\cdot\nabla y_k^{n+1}\cdot w = \langle f^n, w\rangle_{\boldsymbol{V}'\times\boldsymbol{V}} \quad \forall w\in\boldsymbol{V},\ \forall n\geq 0, \\ v_k^0 = 0. \end{cases} \tag{11.100}$$

(ii) Then, to compute the term $\|v_{k,t}\|_{L^2(0,T;\boldsymbol{V}')}$ of $E(y_k)$, introduction of the solution $w_k \in L^2(\boldsymbol{V})$ of

$$\int_0^T\!\!\int_\Omega \nabla w_k\cdot\nabla w + v_{k,t}\cdot w = 0 \quad \forall w\in L^2(\boldsymbol{V}), \tag{11.101}$$

so that $\|v_{k,t}\|_{L^2(0,T;\boldsymbol{V}')} = \|\nabla w_k\|_{L^2(Q_T)}$. An approximation of $w_k$ is obtained through the scheme

$$\int_\Omega \nabla w_k^n\cdot\nabla w + \frac{v_k^{n+1}-v_k^n}{\delta t}\cdot w = 0 \quad \forall w\in\boldsymbol{V}\ \forall n\in\mathbb{N}. \tag{11.102}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

(iii) Computation of an approximation of solution $Y_{1,k}$ of (11.84) through the scheme

$$\begin{cases} \displaystyle\int_\Omega \frac{Y_{1,k}^{n+1}-Y_{1,k}^{n}}{\delta t}\cdot w+\nu\int_\Omega \nabla Y_{1,k}^{n+1}\cdot\nabla w+\int_\Omega y_k^{n+1}\cdot\nabla Y_{1,k}^{n+1}\cdot w \\ \quad+\displaystyle\int_\Omega Y_{1,k}^{n+1}\cdot\nabla y_k^{n+1}\cdot w=-\int_\Omega \frac{v_k^{n+1}-v_k^{n}}{\delta t}\cdot w-\int_\Omega \nabla v_k^{n+1}\cdot\nabla w \\ \quad \forall w\in \boldsymbol{V},\ \forall n\geq 0. \\ Y_{1,k}^{0}=0. \end{cases} \tag{11.103}$$

(iv) Computation of the corrector function $\overline{\overline{v}}_k$, the solution of (11.88), through the scheme

$$\begin{cases} \displaystyle\int_\Omega \frac{\overline{\overline{v}}_k^{n+1}-\overline{\overline{v}}_k^{n}}{\delta t}\cdot w+\int_\Omega \nabla\overline{\overline{v}}_k^{n+1}\cdot\nabla w+\int_\Omega Y_{1,k}^{n+1}\cdot\nabla Y_{1,k}^{n+1}\cdot w=0 \quad \forall w\in\boldsymbol{V},\ n\geq 0, \\ \overline{\overline{v}}_k(0)=0. \end{cases} \tag{11.104}$$

(v) Computation of $\|v_k\|_{\mathcal{A}_0}^2$, $\langle v_k,\overline{\overline{v}}_k\rangle_{\mathcal{A}_0}$, and $\|\overline{\overline{v}}_k\|_{\mathcal{A}_0}^2$ appearing in $E(y_k-\lambda Y_{1,k})$ (see (11.89)). The computation of $\|\overline{\overline{v}}_k\|_{\mathcal{A}_0}$ requires the computation of $\|\overline{\overline{v}}_k\|_{L^2(\boldsymbol{V}')}$, i. e., the introduction of $\overline{\overline{w}}_k$ solution of

$$\int_0^T\int_\Omega \nabla\overline{\overline{w}}_k\cdot\nabla w+v_{k,t}\cdot w=0 \quad \forall w\in L^2(\boldsymbol{V}),$$

so that $\|\overline{\overline{v}}_{k,t}\|_{L^2(\boldsymbol{V}')}=\|\nabla\overline{\overline{w}}_k\|_{L^2(Q_T)}$ through the scheme

$$\int_\Omega \nabla\overline{\overline{w}}_k^{n}\cdot\nabla w+\frac{\overline{\overline{v}}_k^{n+1}-\overline{\overline{v}}_k^{n}}{\delta t}\cdot w=0 \quad \forall w\in\boldsymbol{V},\ \forall n\in\mathbb{N}. \tag{11.105}$$

(vi) Determination of the minimum $\lambda_k\in(0,m]$ of

$$\lambda\to E(y_k-\lambda Y_{1,k})=(1-\lambda)^2\|v_k\|_{\mathcal{A}_0}^2+2\lambda^2(1-\lambda)\langle v_k,\overline{\overline{v}}_k\rangle_{\mathcal{A}_0}+\lambda^4\|\overline{\overline{v}}_k\|_{\mathcal{A}_0}^2$$

by the Newton–Raphson method starting from 0 and, finally, updating the sequence $y_{k+1}=y_k-\lambda_k Y_{1,k}$.

As a summary, the determination of $y_{k+1}$ from $y_k$ involves the resolution of four Stokes-type problems, namely (11.100), (11.102), (11.104), and (11.105) plus the resolution of the linearized Navier–Stokes problem (11.103). This latter concentrates most of

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

the computational time resources since the operator (to be inverted) varies with the index $n$.

Instead of minimizing exactly the fourth-order polynomial $\lambda \to E(y_k - \lambda Y_{1,k})$ in step (vi), we may simply minimize with respect to $\lambda \in (0,1]$ the right-hand side of the estimate

$$E(y_k - \lambda Y_{1,k}) \le \left( |1-\lambda| \sqrt{E(y_k)} + \frac{\lambda^2}{\sqrt{2}} \|\overline{\overline{v}}_k\|_{\mathcal{A}_0} \right)^2$$

(appearing in the proof of Lemma 11.3.15), leading to $\widehat{\lambda}_k = \min(1, \frac{\sqrt{E(y_k)}}{\sqrt{2}\|\overline{\overline{v}}_k\|_{\mathcal{A}_0}})$ (see Remark 11.3.21). This avoids the computation of the scalar product $\langle v_k, \overline{\overline{v}}_k \rangle_{\mathcal{A}_0}$ and solving one of Stokes-type problems.

**Remark 11.4.1.** Similarly, we may also consider the equivalent functional $\widetilde{E}$ defined in (11.10). This avoids the introduction of the auxiliary corrector function $v$ and reduces to three (instead of four) the number of Stokes-type problems to be solved. Precisely, using the initialization defined in (11.98), the algorithm is as follows:

(i) Computation of $\widetilde{E}(y_k) = \|h_k\|_{L^2(\boldsymbol{V})} = \|\nabla h_k\|_{L^2(Q_T)}$, where $h_k$ solves

$$\int_0^T \int_\Omega \nabla h_k \cdot \nabla w + (y_{k,t} - \nu \Delta y_k + y_k \cdot \nabla y_k - f) \cdot w = 0 \quad \forall w \in L^2(\boldsymbol{V})$$

through the scheme

$$\begin{aligned} &\int_\Omega \nabla h_k^n \cdot \nabla w + \frac{y_k^{n+1} - y_k^n}{\delta t} \cdot w + \nu \nabla y_k^{n+1} \cdot \nabla w + y_k^{n+1} \cdot \nabla y_k^{n+1} \\ &\quad = \langle f^n, w \rangle_{\boldsymbol{V}', \boldsymbol{V}} \quad \forall w \in \boldsymbol{V},\ \forall n \in \mathbb{N}. \end{aligned} \tag{11.106}$$

(ii) Computation of an approximation of $Y_{1,k}$ from $y_k$ through the scheme

$$\begin{cases} \displaystyle\int_\Omega \frac{Y_{1,k}^{n+1} - Y_{1,k}^n}{\delta t} \cdot w + \nu \int_\Omega \nabla Y_{1,k}^{n+1} \cdot \nabla w + \int_\Omega y_k^{n+1} \cdot \nabla Y_{1,k}^{n+1} \cdot w + \int_\Omega Y_k^{n+1} \cdot \nabla y_{1,k}^{n+1} \cdot w \\ \quad = \displaystyle\int_\Omega \frac{y_k^{n+1} - y_k^n}{\delta t} \cdot w + \nu \int_\Omega \nabla y_k^{n+1} \cdot \nabla w \\ \quad\quad + \displaystyle\int_\Omega y_k^{n+1} \cdot \nabla y_k^{n+1} \cdot w - \langle f^n, w \rangle_{\boldsymbol{V}' \times \boldsymbol{V}} \quad \forall w \in \boldsymbol{V},\ \forall n \ge 0, \\ Y_{1,k}^0 = 0. \end{cases} \tag{11.107}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

(iii) Computation of $\|B(Y_{1,k}, Y_{1,k})\|_{L^2(0,T;\boldsymbol{V}')} = \|h_k\|_{L^2(\boldsymbol{V})} = \|\nabla h_k\|_{L^2(Q_T)}$, where $h_k$ solves

$$\int_0^T \iint_\Omega \nabla h_k \cdot \nabla w + Y_{1,k} \cdot \nabla Y_{1,k} \cdot w = 0 \quad \forall w \in L^2(\boldsymbol{V}),$$

and similarly of the term $\langle y_{k,t} + \nu B_1(y_k) + B(y_k, y_k), B(Y_{1,k}, Y_{1,k})\rangle_{L^2(0,T;\boldsymbol{V}')}$.

(iv) Determination of the minimum $\lambda_k \in (0, m]$ of

$$\begin{aligned}\lambda \to \widetilde{E}(y_k - \lambda Y_{1,k}) = (1-\lambda)^2 \widetilde{E}(y_k) &\\ + \lambda^2(1-\lambda)\big\langle y_{k,t} + \nu B_1(y_k) + B(y_k, y_k) - f, B(Y_{1,k}, Y_{1,k})\big\rangle_{L^2(0,T;\boldsymbol{V}')} &\\ + \frac{\lambda^4}{2}\big\|B(Y_{1,k}, Y_{1,k})\big\|^2_{L^2(0,T;\boldsymbol{V}')} &\end{aligned}$$

by the Newton–Raphson method starting from 0 and finally update of the sequence $y_{k+1} = y_k - \lambda_k Y_{1,k}$ until $\widetilde{E}(y_k)$ is small enough.

We emphasize once more that the case $\lambda_k = 1$ coincides with the standard Newton algorithm to find zeros of the functional $F : \mathcal{A} \to L^2(0, T; \boldsymbol{V}')$ defined by $F(y) = y_t + \nu B_1(y) + B(y, y) - f$. In terms of computational time resources, the determination of the optimal descent step $\lambda_k$ is negligible with respect to the resolution in step (ii).

## 11.4.2 2D semicircular driven cavity

We illustrate our theoretical results for the 2D semicircular cavity discussed in [9]. The geometry is a semidisk $\Omega = \{(x_1, x_2) \in \mathbb{R}^2, x_1^2 + x_2^2 < 1/4, x_2 \le 0\}$ depicted in Figure 11.1. The velocity is imposed to $y = (g, 0)$ on $\Gamma_0 = \{(x_1, 0) \in \mathbb{R}^2, |x_1| < 1/2\}$ with $g$ vanishing at $x_1 = \pm 1/2$ and close to one elsewhere: we take $g(x_1) = (1 - e^{100(x_1 - 1/2)})(1 - e^{-100(x_1 + 1/2)})$. On the complement $\Gamma_1 = \{(x_1, x_2) \in \mathbb{R}^2, x_2 < 0, x_1^2 + x_2^2 = 1/4\}$ of the boundary the velocity is fixed to zero.

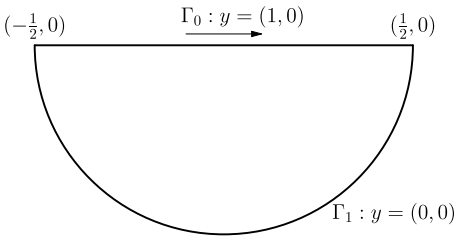



**Figure 11.1:** Semidisk geometry.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

This example has been used in [14] to solve the corresponding steady problem (for which the weak solution is not unique), using again an iterative least-squares strategy. There the method proved to be robust enough for small values of $\nu$ of order $10^{-4}$, whereas the standard Newton method failed. Figure 11.2 depicts the streamlines of steady-state solutions corresponding to $\nu^{-1} = 500$ and to $\nu^{-1} = i{\times}10^3$ for $i = 1, \ldots, 9$. The values used to plot the stream function are given in Table 11.1. The figures are in very good agreements with those depicted in [9]. When the Reynolds number (here equal to $\nu^{-1}$) is small, the final steady state consists of one vortex. As the Reynolds number in-

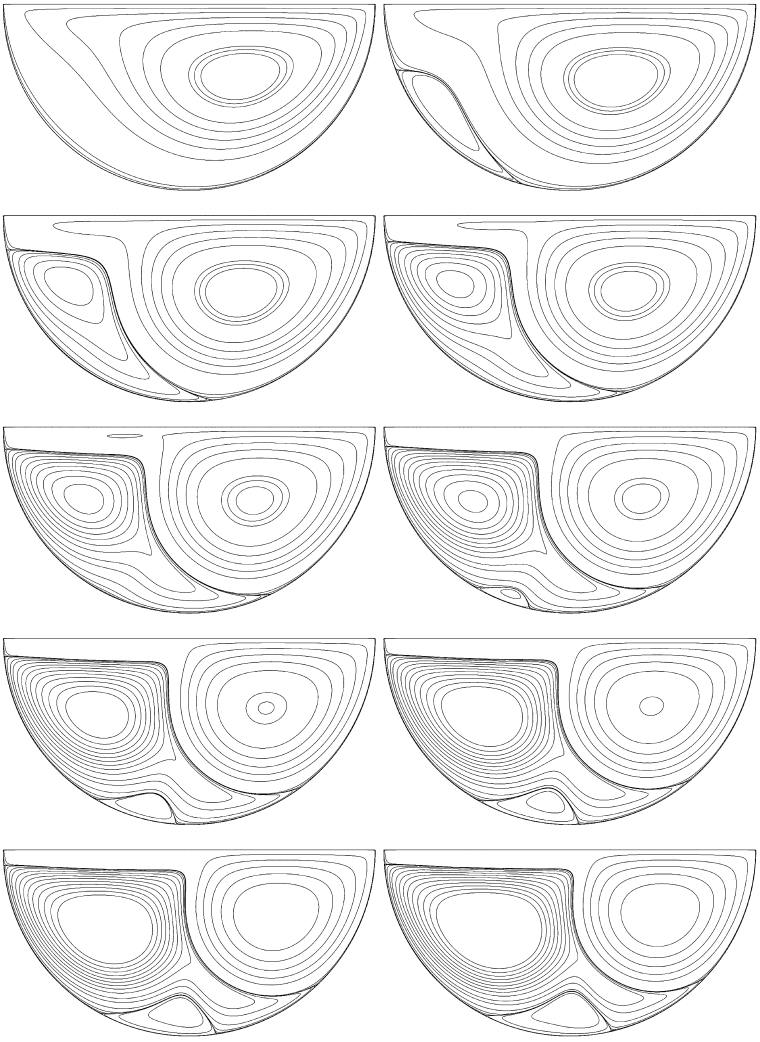


**Figure 11.2:** Streamlines of the steady-state solution for $\nu^{-1} = 500$, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, and 9000.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Table 11.1:** Values used to plot the contours of the stream function.

| |
|---|
| $-0.07, -0.0675, -0.065, -0.05, -0.04, -0.03, -0.02, -0.01, \pm 10^{-4}, \pm 10^{-5}$ $\pm 10^{-7}, -10^{-10}, 0, 10^{-8}, 10^{-6}, 5 \times 10^{-4}, 10^{-3}, 2 \times 10^{-3}, 3 \times 10^{-3}, 4 \times 10^{-3}$ $5 \times 10^{-3}, 6 \times 10^{-3}, 7 \times 10^{-3}, 8 \times 10^{-3}, 9 \times 10^{-3}, 0.01$ |

creases, first, a secondary vortex and then a tertiary vortex arises, whose size depends on the Reynolds number too. Moreover, according to [9], when the Reynolds number exceeds approximatively 6 650, a Hopf bifurcation phenomenon occurs in the sense that the unsteady solution does not reach a steady state anymore (as time evolves) but shows an oscillatory behavior. We mention that the Navier–Stokes system is solved in [9] using an operator-splitting/finite element-based methodology. In particular, concerning the time discretization, an explicit scheme is employed.

## 11.4.3 Experiments

We report some numerical results performed with the FreeFem++ package developed at Sorbonne university (see [11]). Regular triangular meshes are used together with the $\mathbb{P}_2/\mathbb{P}_1$ Taylor–Hood finite element, satisfying the Ladyzhenskaya–Babushka–Brezzi condition of stability. An example of mesh composed of 9 063 triangles is displayed in Figure 11.3.

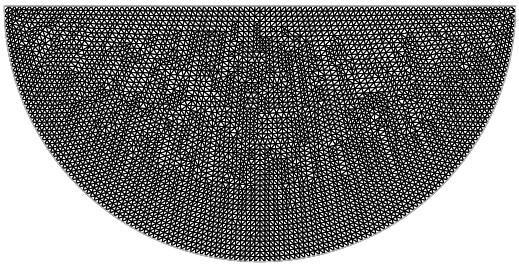


**Figure 11.3:** A regular triangulation of the semidisk geometry; ♯triangles = 9 064; ♯vertices = 4 663; size $h \approx 1.62 \times 10^{-2}$.

To deeply emphasize the influence of the value of $\nu$ on the behavior of the algorithm described in Section 11.4.1, we consider an initial guess $y_0$ of the sequence $(y_k)_{(k>0)}$ independent of $\nu$. Precisely, we define $y_0$ as the solution of the unsteady Stokes system with viscosity equal to one (i. e., $\overline{\nu} = 1$ in (11.98)) and source term $f \equiv 0$. The initial condition $u_0 \in \boldsymbol{H}$ is defined as the solution of $-\Delta u_0 + \nabla p = 0$, $\nabla \cdot u_0 = 0$ in $\Omega$ and boundary conditions $u_0 = g$ on $\Gamma_0$ and $u_0 = 0$ on $\Gamma_1$; $u_0$ in fact belongs to $\boldsymbol{V}$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Tables 11.2 and 11.3 report numerical values of the sequences $(\sqrt{2E(y_k)})_{(k>0)}$, $(\lambda_k)_{(k>0)}$, and $(\|y_k - y_{k-1}\|_{L^2(\boldsymbol{V})}/\|y_k\|_{L^2(\boldsymbol{V})})_{(k>0)}$ associated with $\nu = 1/500$ and $\nu = 1/1000$, respectively, and $T = 10$ and $f = 0$. The tables also display (on the right part) the values obtained when the parameter $\lambda_k$ is a fixed constant equal to one, corresponding to the standard Newton method. The algorithms are stopped when $\sqrt{2E(y_k)} \leq 10^{-8}$. The triangular mesh of Figure 11.3 for which the discretization parameter $h$ equals $1.62 \times 10^{-2}$ is employed. The number of degrees of freedom is 23 315. Moreover, the time discretization parameter in $\delta t$ is taken equal to $10^{-2}$.

**Table 11.2:** $\nu = 1/500$; Results for algorithm (11.29).

| ♯iterate $k$ | $\frac{\|y_k - y_{k-1}\|_{L^2(V)}}{\|y_{k-1}\|_{L^2(V)}}$ | $\sqrt{2E(y_k)}$ | $\lambda_k$ | $\frac{\|y_k - y_{k-1}\|_{L^2(V)}}{\|y_{k-1}\|_{L^2(V)}}$ $(\lambda_k = 1)$ | $\sqrt{2E(y_k)}$ $(\lambda_k = 1)$ |
|---|---|---|---|---|---|
| 0 | – | $2.690 \times 10^{-2}$ | 0.8112 | – | $2.690 \times 10^{-2}$ |
| 1 | $4.540 \times 10^{-1}$ | $1.077 \times 10^{-2}$ | 0.7758 | $5.597 \times 10^{-1}$ | $1.254 \times 10^{-2}$ |
| 2 | $1.836 \times 10^{-1}$ | $3.653 \times 10^{-3}$ | 0.8749 | $2.236 \times 10^{-1}$ | $5.174 \times 10^{-3}$ |
| 3 | $7.503 \times 10^{-2}$ | $7.794 \times 10^{-4}$ | 0.9919 | $7.830 \times 10^{-2}$ | $6.133 \times 10^{-4}$ |
| 4 | $1.437 \times 10^{-2}$ | $2.564 \times 10^{-5}$ | 1.0006 | $9.403 \times 10^{-3}$ | $1.253 \times 10^{-5}$ |
| 5 | $4.296 \times 10^{-4}$ | $3.180 \times 10^{-8}$ | 1 | $1.681 \times 10^{-4}$ | $4.424 \times 10^{-9}$ |
| 6 | $5.630 \times 10^{-7}$ | $6.384 \times 10^{-11}$ | – | – | – |

**Table 11.3:** $\nu = 1/1000$; Results for algorithm (11.83).

| ♯iterate $k$ | $\frac{\|y_k - y_{k-1}\|_{L^2(V)}}{\|y_{k-1}\|_{L^2(V)}}$ | $\sqrt{2E(y_k)}$ | $\lambda_k$ | $\frac{\|y_k - y_{k-1}\|_{L^2(V)}}{\|y_{k-1}\|_{L^2(V)}}$ $(\lambda_k = 1)$ | $\sqrt{2E(y_k)}$ $(\lambda_k = 1)$ |
|---|---|---|---|---|---|
| 0 | – | $2.69 \times 10^{-2}$ | 0.634 | – | $2.69 \times 10^{-2}$ |
| 1 | $5.13 \times 10^{-1}$ | $1.49 \times 10^{-2}$ | 0.580 | $8.10 \times 10^{-1}$ | $2.23 \times 10^{-2}$ |
| 2 | $2.53 \times 10^{-1}$ | $7.60 \times 10^{-3}$ | 0.349 | $4.45 \times 10^{-1}$ | $2.91 \times 10^{-2}$ |
| 3 | $1.34 \times 10^{-1}$ | $5.47 \times 10^{-3}$ | 0.402 | $5.71 \times 10^{-1}$ | $5.68 \times 10^{-2}$ |
| 4 | $1.10 \times 10^{-1}$ | $3.81 \times 10^{-3}$ | 0.561 | $3.68 \times 10^{-1}$ | $2.62 \times 10^{-2}$ |
| 5 | $8.95 \times 10^{-2}$ | $2.29 \times 10^{-3}$ | 0.868 | $2.86 \times 10^{-1}$ | $1.82 \times 10^{-2}$ |
| 6 | $6.39 \times 10^{-2}$ | $8.67 \times 10^{-4}$ | 1.036 | $1.42 \times 10^{-1}$ | $4.30 \times 10^{-3}$ |
| 7 | $1.78 \times 10^{-2}$ | $4.15 \times 10^{-5}$ | 0.999 | $6.05 \times 10^{-2}$ | $9.60 \times 10^{-4}$ |
| 8 | $7.98 \times 10^{-4}$ | $9.93 \times 10^{-8}$ | 0.999 | $1.48 \times 10^{-2}$ | $5.66 \times 10^{-5}$ |
| 9 | $2.25 \times 10^{-6}$ | $4.00 \times 10^{-11}$ | – | $9.74 \times 10^{-4}$ | $3.02 \times 10^{-7}$ |
| 10 | – | – | – | $4.26 \times 10^{-6}$ | $3.84 \times 10^{-11}$ |

For $\nu = 1/500$, the optimal $\lambda_k$ are close to one ($\max_k |1 - \lambda_k| \leq 1/5$), so that the two algorithms produce very similar behaviors. The convergence is observed after six iterations. For $\nu = 1/1000$, we observe that the optimal $\lambda_k$ are far from one during the first iterates. The optimization of the parameter allows a faster convergence (after nine iterates) than the usual Newton method. For instance, after eight iterates,

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$\sqrt{2E(y_k)} \approx 9.931 \times 10^{-11}$ in the first case, and $\sqrt{2E(y_k)} \approx 5.669 \times 10^{-5}$ in the second one. In agreement with the theoretical results, we also check that $\lambda_k$ goes to one. Moreover, the decrease of $\sqrt{2E(y_k)}$ is first linear and then (when $\lambda_k$ becomes close to one) quadratic.

At time $T = 10$, the unsteady state solution is close to the solution of the steady Navier–Stokes equation: the last element $y_{k=9}$ of the converged sequence satisfies $\|y_{k=9}(T,\cdot) - y_{k=9}(T-\delta t,\cdot)\|_{L^2(\Omega)}/\|y_{k=9}(T,\cdot)\|_{L^2(\Omega)} \approx 1.19 \times 10^{-5}$. Figure 11.4 displays the streamlines of the unsteady state solution corresponding to $\nu = 1/1000$ at times 0, 1, 2, 3, 4, 5, 6, and 7 seconds to be compared with the streamlines of the steady solution depicted in Figure 11.2.

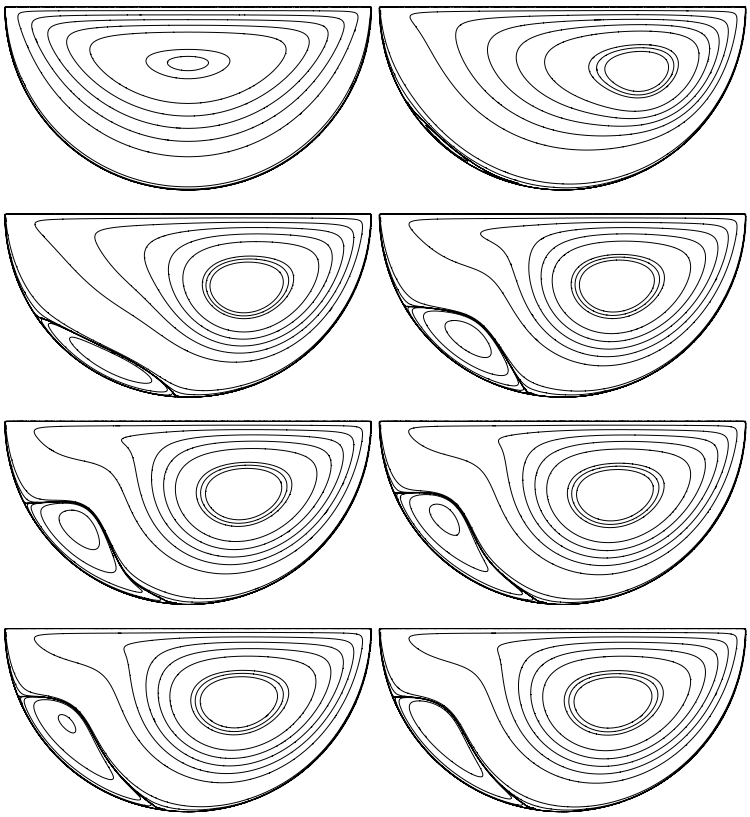


**Figure 11.4:** Streamlines of the unsteady state solution for $\nu^{-1} = 1000$ at time $t = i$, $i = 0, \ldots, 7$ s.

For lower values of the viscosity constant, namely, approximately $\nu \leq 1/1100$, the initial guess $y_0$ is too far from the zero of $E$, so that we observe the divergence after few iterates of the Newton method (case $\lambda_k = 1$ for all $k > 0$) but still the convergence of the algorithm described in Section 11.4.1 (see Table 11.4). The divergence in the case $\lambda_k = 1$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Table 11.4:** $\nu = 1/1100$; Results for algorithm (11.83).

| ♯iterate $k$ | $\frac{\|y_k - y_{k-1}\|_{L^2(V)}}{\|y_{k-1}\|_{L^2(V)}}$ | $\sqrt{2E(y_k)}$ | $\lambda_k$ | $\frac{\|y_k - y_{k-1}\|_{L^2(V)}}{\|y_{k-1}\|_{L^2(V)}}$ $(\lambda_k = 1)$ | $\sqrt{2E(y_k)}$ $(\lambda_k = 1)$ |
|---|---|---|---|---|---|
| 0 | – | $2.69 \times 10^{-2}$ | 0.614 | – | $2.69 \times 10^{-2}$ |
| 1 | $5.24 \times 10^{-1}$ | $1.53 \times 10^{-2}$ | 0.566 | $8.52 \times 10^{-1}$ | $2.38 \times 10^{-2}$ |
| 2 | $2.64 \times 10^{-1}$ | $8.02 \times 10^{-3}$ | 0.323 | $4.89 \times 10^{-1}$ | $3.55 \times 10^{-2}$ |
| 3 | $1.38 \times 10^{-1}$ | $5.98 \times 10^{-3}$ | 0.330 | $7.17 \times 10^{-1}$ | $8.70 \times 10^{-2}$ |
| 4 | $1.11 \times 10^{-1}$ | $4.54 \times 10^{-3}$ | 0.420 | $4.84 \times 10^{-1}$ | $3.53 \times 10^{-2}$ |
| 5 | $9.42 \times 10^{-2}$ | $3.22 \times 10^{-3}$ | 0.587 | $1.12 \times 10^{0}$ | $3.90 \times 10^{-1}$ |
| 6 | $7.66 \times 10^{-2}$ | $1.94 \times 10^{-3}$ | 0.972 | – | $1.33 \times 10^{4}$ |
| 7 | $5.68 \times 10^{-2}$ | $5.93 \times 10^{-4}$ | 1.021 | – | $8.09 \times 10^{27}$ |
| 8 | $1.00 \times 10^{-2}$ | $1.08 \times 10^{-5}$ | 0.999 | – | – |
| 9 | $2.83 \times 10^{-4}$ | $1.33 \times 10^{-8}$ | 1 | – | – |
| 10 | $2.89 \times 10^{-7}$ | $4.61 \times 10^{-11}$ | – | – | – |

is still observed with a refined discretization both in time and space, corresponding to $\delta t = 0.5 \times 10^{-3}$ and $h \approx 0.0110$ (19 810 triangles and 10 101 vertices). The divergence of the Newton method suggests that the functional $E$ is not convex far away from the zero of $E$ and that the derivative $E'(y)$ takes small values there. We recall that, in view of the theoretical part, the functional $E$ is coercive, and its derivative vanishes only at the zero of $E$. However, the equality $E'(y_k) \cdot Y_{1,k} = 2E(y_k)$ shows that $E'(y_k)$ can be "small" for "large" $Y_{1,k}$, i. e., "large" $y_k$. On the other hand, we observe the convergence (after three iterates) of the Newton method, when initialized with the approximation corresponding to $\nu = 1/1000$.

Table 11.5 gives numerical values associated with $\nu = 1/2000$ and $T = 10$. We used a refined discretization: precisely, $\delta t = 1/150$ and a mesh composed of 15 190 triangles, 7 765 vertices ($h \approx 1.343 \times 10^{-2}$). The convergence of the algorithm of Section 11.4.1 is observed after 19 iterates. In agreement with the theoretical results, the sequence $\{\lambda_k\}_{(k>0)}$ goes to one. Moreover, the variation of the error functional $E(y_k)$ is first quite slow and then increases to be very fast after 15 iterates. This behavior is illustrated in Figure 11.5. For lower values of $\nu$, we still observed the convergence (provided a fine enough discretization so as to capture the third vortex) with an increasing number of iterates. For instance, 28 iterates are necessary to achieve $\sqrt{2E(y_k)} \leq 10^{-8}$ for $\nu = 1/3000$ and 49 iterates for $\nu = 1/4000$. This illustrates the global convergence of the algorithm. In view of estimate (11.92), a quadratic rate is achieved as soon as $\sqrt{E(y_k)} \leq C_1^{-1}$ with (since $f \equiv 0$)

$$C_1 = \frac{c}{\nu\sqrt{\nu}} \exp\left(\frac{c}{\nu^2}\|u_0\|_{\boldsymbol{H}}^2 + \frac{c}{\nu^3}E(y_0)\right),$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Table 11.5:** $\nu = 1/2000$; Results for algorithm (11.83).

| ♯iterate $k$ | $\frac{\|y_k - y_{k-1}\|_{L^2(V)}}{\|y_{k-1}\|_{L^2(V)}}$ | $\sqrt{2E(y_k)}$ | $\lambda_k$ |
|---|---|---|---|
| 0 | – | $2.691 \times 10^{-2}$ | 0.5215 |
| 1 | $6.003 \times 10^{-1}$ | $1.666 \times 10^{-2}$ | 0.4919 |
| 2 | $3.292 \times 10^{-1}$ | $9.800 \times 10^{-3}$ | 0.1566 |
| 3 | $1.375 \times 10^{-1}$ | $8.753 \times 10^{-3}$ | 0.1467 |
| 4 | $1.346 \times 10^{-1}$ | $7.851 \times 10^{-3}$ | 0.0337 |
| 5 | $5.851 \times 10^{-2}$ | $7.688 \times 10^{-3}$ | 0.0591 |
| 6 | $7.006 \times 10^{-2}$ | $7.417 \times 10^{-3}$ | 0.1196 |
| 7 | $9.691 \times 10^{-2}$ | $6.864 \times 10^{-3}$ | 0.0977 |
| 8 | $8.093 \times 10^{-2}$ | $6.465 \times 10^{-3}$ | 0.0759 |
| 9 | $6.400 \times 10^{-2}$ | $6.182 \times 10^{-3}$ | 0.0968 |
| 10 | $6.723 \times 10^{-2}$ | $5.805 \times 10^{-3}$ | 0.1184 |
| 11 | $6.919 \times 10^{-2}$ | $5.371 \times 10^{-3}$ | 0.1630 |
| 12 | $7.414 \times 10^{-2}$ | $4.825 \times 10^{-3}$ | 0.2479 |
| 13 | $8.228 \times 10^{-2}$ | $4.083 \times 10^{-3}$ | 0.3517 |
| 14 | $8.146 \times 10^{-2}$ | $3.164 \times 10^{-3}$ | 0.4746 |
| 15 | $7.349 \times 10^{-2}$ | $2.207 \times 10^{-3}$ | 0.7294 |
| 16 | $6.683 \times 10^{-2}$ | $1.174 \times 10^{-3}$ | 1.0674 |
| 17 | $3.846 \times 10^{-2}$ | $2.191 \times 10^{-4}$ | 1.0039 |
| 18 | $5.850 \times 10^{-3}$ | $4.674 \times 10^{-5}$ | 0.9998 |
| 19 | $1.573 \times 10^{-4}$ | $5.843 \times 10^{-9}$ | – |

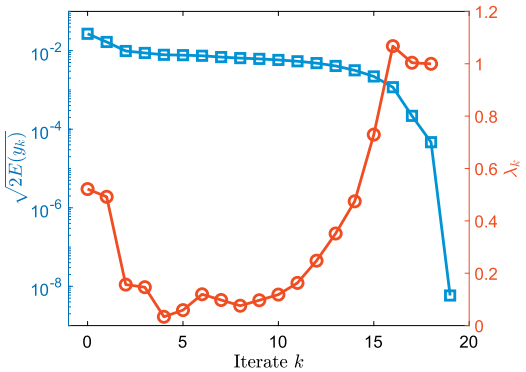



**Figure 11.5:** Evolution of $\sqrt{2E(y_k)}$ and $\lambda_k$ with respect to $k$; $\nu = 1/2000$ (see Table 11.5).

so that $C_1^{-1} \to 0$ as $\nu \to 0$. Consequently, for small $\nu$, it is very likely more efficient (in terms of computational resources) to couple the algorithm with a continuation method with respect to $\nu$ to reach faster the quadratic regime. This aspect is not addressed in this work, and we refer to [14], where this is illustrated in the steady case.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

## 11.5 Conclusions and perspectives

In order to get an approximation of the solutions of the steady and unsteady 2D Navier–Stokes equations, we have introduced and analyzed a least-squares method based on a minimization of an appropriate norm of the equation. For instance, in the unsteady case, considering the weak solution associated with an initial condition in $\boldsymbol{H} \subset L^2(\Omega)^2$ and a source $f \in L^2(0,T;H^{-1}(\Omega)^2)$, the least-squares functional is based on the $L^2(0,T;\boldsymbol{V'})$-norm of the state equation. Using a particular descent direction, we construct explicitly a minimizing sequence for the functional converging strongly, for any initial guess, to the solution of the Navier–Stokes equation. Moreover, except for the first iterates, the convergence is quadratic. In fact, it turns out that this minimizing sequence coincides with the sequence obtained from the damped Newton method when used to solve the weak problem associated with the Navier–Stokes equation. The numerical experiments illustrate the global convergence of the method and its robustness including for small values of the viscosity constant.

Moreover, the strong convergence of the whole minimizing sequence has been proved using a coercivity-type property of the functional, a consequence of the uniqueness of the solution. In fact, it is interesting to remark that this property is not necessary, since such minimizing sequence (which is completely determined by the initial term) is a Cauchy sequence. The approach can therefore be adapted to partial differential equations with multiple solutions or to optimization problem involving various solutions. We mention notably the approximation of null controls for (controllable) nonlinear partial differential equations: the source term $f$, possibly distributed over a nonempty set of $\Omega$ is now, together with the corresponding state, an argument of the least-squares functional. The controllability constraint is incorporated in the set $\mathcal{A}$ of admissible pair $(y,f)$. In spite of the nonuniqueness of the minimizers, the approach introduced in this work may be used to construct a strongly convergent approximation of null controls. We refer to [12] for the analysis of this approach for null controllable semilinear heat equation.

## Bibliography

[1] P. B. Bochev and M. D. Gunzburger. *Least-Squares Finite Element Methods*, volume 166 of Applied Mathematical Sciences. Springer, New York, 2009.

[2] M. O. Bristeau, O. Pironneau, R. Glowinski, J. Periaux, and P. Perrier. On the numerical solution of nonlinear problems in fluid dynamics by least squares and finite element methods. I. Least square formulations and conjugate gradient solution of the continuous problems. *Comput. Methods Appl. Mech. Eng.*, 17/18(part 3):619–657, 1979.

[3] R. Coifman, P.-L. Lions, Y. Meyer, and S. Semmes. Compensated compactness and Hardy spaces. *J. Math. Pures Appl. (9)*, 72(3):247–286, 1993.

[4] P. Deuflhard. *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms*, volume 35 of Springer Series in Computational Mathematics. Springer, Heidelberg, 2011. First softcover printing of the 2006 corrected printing.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[5] M. Farazmand. An adjoint-based approach for finding invariant solutions of Navier–Stokes equations. *J. Fluid Mech.*, 795:278–312, 2016.
[6] V. Girault and P.-A. Raviart. *Finite Element Methods for Navier–Stokes Equations: Theory and Algorithms*, volume 5 of Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 1986.
[7] R. Glowinski. Finite element methods for incompressible viscous flow. In *Numerical Methods for Fluids (Part 3)*, volume 9 of Handbook of Numerical Analysis, pages 3–1176. Elsevier, 2003.
[8] R. Glowinski. *Variational Methods for the Numerical Solution of Nonlinear Elliptic Problems*, volume 86 of CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2015.
[9] R. Glowinski, G. Guidoboni, and T.-W. Pan. Wall-driven incompressible viscous flow in a two-dimensional semi-circular cavity. *J. Comput. Phys.*, 216(1):76–91, 2006.
[10] R. Glowinski, B. Mantel, J. Periaux, and O. Pironneau. $H^{-1}$ least squares method for the Navier–Stokes equations. In *Numerical Methods in Laminar and Turbulent Flow (Proc. First Internat. Conf., Univ. College Swansea, Swansea, 1978)*, pages 29–42. Halsted, New York–Toronto, Ont., 1978.
[11] F. Hecht. New development in Freefem++. *J. Numer. Math.*, 20(3–4):251–265, 2012.
[12] J. Lemoine, I. Marin-Gayte, and A. Münch. Approximation of null controls for semilinear heat equations using a least-squares approach. *ESAIM COCV*, 27(63):1–28, 2021.
[13] J. Lemoine and A. Münch. A fully space-time least-squares method for the unsteady Navier–Stokes system. *J. Math. Fluid Mech.*, 23(4):1–102, 2021.
[14] J. Lemoine and A. Münch. Resolution of the implicit Euler scheme for the Navier–Stokes system through a least-squares method. *J. Numer. Math.*, 147(2):349–391, 2021.
[15] J. Lemoine, A. Münch, and P. Pedregal. Analysis of continuous $H^{-1}$-least-squares approaches for the steady Navier–Stokes system. *Appl. Math. Optim.*, 83(1):461–488, 2021.
[16] J.-L. Lions. *Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires*. Dunod; Gauthier-Villars, Paris, 1969.
[17] M. Ortiz, B. Schmidt, and U. Stefanelli. A variational approach to Navier–Stokes. *Nonlinearity*, 31(12):5664–5682, 2018.
[18] P. Pedregal. A variational approach for the Navier–Stokes system. *J. Math. Fluid Mech.*, 14(1):159–176, 2012.
[19] O. Pironneau. *Finite Element Methods for Fluids*. John Wiley & Sons, Ltd., Chichester; Masson, Paris, 1989. Translated from the French.
[20] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*, volume 23 of Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 1994.
[21] P. Saramito. A damped Newton algorithm for computing viscoplastic fluid flows. *J. Non-Newton. Fluid Mech.*, 238:6–15, 2016.
[22] J. Simon. Nonhomogeneous viscous incompressible fluids: existence of velocity, density, and pressure. *SIAM J. Math. Anal.*, 21(5):1093–1117, 1990.
[23] A. Smith and D. Silvester. Implicit algorithms and their linearization for the transient incompressible Navier–Stokes equations. *IMA J. Numer. Anal.*, 17(4):527–545, 1997.
[24] L. Tartar. *An Introduction to Navier–Stokes Equation and Oceanography*, volume 1 of Lecture Notes of the Unione Matematica Italiana. Springer-Verlag, Berlin; UMI, Bologna, 2006.
[25] R. Temam. *Navier–Stokes Equations: Theory and Numerical Analysis*. AMS Chelsea Publishing, Providence, RI, 2001. Reprint of the 1984 edition.
[26] F. Tone and D. Wirosoetisno. On the long-time stability of the implicit Euler scheme for the two-dimensional Navier–Stokes equations. *SIAM J. Numer. Anal.*, 44(1):29–40, 2006.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Gontran Lance, Emmanuel Trélat, and Enrique Zuazua

# 12 Numerical issues and turnpike phenomenon in optimal shape design

**Abstract:** This paper follows and complements [12], where we have established the turnpike property for some optimal shape design problems.

Considering linear parabolic partial differential equations where the shapes to be optimized act as a source term, we want to minimize a quadratic criterion. The existence of optimal shapes is proved under some appropriate assumptions. We prove and provide numerical evidence of the turnpike phenomenon for those optimal shapes, meaning that the extremal time-varying optimal solution remains essentially stationary; in fact, it remains essentially close to the optimal solution of an associated static problem.

**Keywords:** optimal shape design, turnpike, numerical analysis, optimal control

**MSC 2010:** 49Q10, 49M41

## Introduction

This paper is devoted to studying large-time shape design problems and to providing evidence of the turnpike phenomenon, stating that the optimal time-varying design mostly stays stationary along the time interval. We complete and comment on results in [12], and we focus on the numerical solving of those large-time time-varying shape

**Acknowledgement:** This project has received funding from the Grants ICON-ANR-16-ACHN-0014 and Finite4SoS ANR-15-CE23-0007-01 of the French ANR, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 694126-DyCon), the Alexander von Humboldt-Professorship program, the Air Force Office of Scientific Research under Award NO: FA9550-18-1-0242, Grant MTM2017-92996-C2-1-R COSNET of MINECO (Spain) and by the ELKARTEK project KK-2018/00083 ROAD2DC of the Basque Government, Transregio 154 Project *Mathematical Modelling, Simulation and Optimization using the Example of Gas Networks* of the German DFG, the European Union Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement 765579-ConFlex.

**Gontran Lance, Emmanuel Trélat,** Sorbonne Université, CNRS, Université de Paris, Inria, Laboratoire Jacques-Louis Lions (LJLL), F-75005 Paris, France, e-mails: lance@ljll.math.upmc.fr, emmanuel.trelat@sorbonne-universite.fr

**Enrique Zuazua,** Chair in Applied Analysis, Alexander von Humboldt-Professorship, Department of Mathematics, Friedrich-Alexander-Universität, Erlangen-Nürnberg, 91058 Erlangen, Germany; and Chair of Computational Mathematics, Fundación Deusto, Av. de las Universidades 24, 48007 Bilbao, Basque Country, Spain; and Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Madrid, Spain, e-mail: enrique.zuazua@fau.de

https://doi.org/10.1515/9783110695984-012

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

design problems that we illustrate with a number of numerical simulations. In Section 12.1, we introduce a general framework and theoretical results, which are then numerically illustrated along the paper. We use the relaxation method, which is classical in optimal shape design. Section 12.2 focuses on the numerical implementation issues, with an approach following the theoretical study. Some numerical illustrations of the turnpike phenomenon in shape design are given in Section 12.3. In Section 12.4, we give some open questions and issues. We illustrate some geometric properties of optimal solutions by numerical examples. A section is particularly set apart for the open issue of the kinetic interpretation of shallow water equations for solving a given shape design problem. We finally provide some proofs in an appendix.

# 12.1 Optimal shape design in parabolic case

## 12.1.1 Context

We denote by:
- $|Q|$ the Lebesgue measure of a measurable subset $Q \subset \mathbf{R}^d$, $d \geq 1$;
- $(p, q)$ the scalar product in $L^2(\Omega)$ of $p$, $q$ in $L^2(\Omega)$;
- $\|y\|$ the $L^2$-norm of $y \in L^2(\Omega)$;
- $\chi_\omega$ the indicator (or characteristic) function of $\omega \subset \mathbf{R}^d$;
- $d_\omega$ the distance function to the set $\omega \subset \mathbf{R}^d$.

Let $\Omega \subset \mathbf{R}^d$ ($d \geq 1$) be an open bounded Lipschitz domain. We consider the uniformly elliptic second-order differential operator

$$Ay = -\sum_{i,j=1}^{d} \partial_{x_j}(a_{ij}(x)\partial_{x_i} y) + \sum_{i=1}^{d} b_i(x)\partial_{x_i} y + c(x)y$$

with $a_{ij}, b_i \in C^1(\Omega)$ and $c \in L^\infty(\Omega)$ with $c \geq 0$. We consider the operator $(A, D(A))$ defined on the domain $D(A)$ encoding Dirichlet conditions $y_{|\partial\Omega} = 0$; when $\Omega$ is $C^2$ or a convex polytop in $\mathbf{R}^2$, we have $D(A) = H_0^1(\Omega) \cap H^2(\Omega)$. The adjoint operator $A^*$ of $A$, also defined on $D(A)$ with homogeneous Dirichlet conditions, is given by

$$A^* v = -\sum_{i,j=1}^{d} \partial_{x_i}(a_{ij}(x)\partial_{x_j} v) - \sum_{i=1}^{d} b_i(x)\partial_{x_i} v + \left(c - \sum_{i=1}^{d} \partial_{x_i} b_i\right) v$$

and is also uniformly elliptic; see [8, Chapter 6]. The operators $A$ and $A^*$ do not depend on $t$ and have a constant of ellipticity $\theta > 0$ (for $A$ written in the *nondivergence form*),

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

i. e.,

$$\sum_{i,j=1}^{d} a_{ij}(x)\xi_i\xi_j \geq \theta|\xi|^2 \quad \forall x \in \Omega.$$

Moreover, we assume that

$$\theta > \theta_1, \tag{12.1}$$

where $\theta_1$ is the largest root of the polynomial

$$P(X) = \frac{X^2}{4\min(1, C_p)} - \|c\|_{L^\infty(\Omega)}X - \frac{\sum_{i=1}^{d}\|b_i\|_{L^\infty(\Omega)}}{2},$$

where $C_p$ is the Poincaré constant of $\Omega$. This assumption is used to ensure that an energy inequality is satisfied with constants not depending on the final time $T$ (see Appendix 12.A.2).

We assume throughout that the operator $A$ satisfies the classical maximum principle (see [8, Section 6.4]) and that

$$c^* = c - \sum_{i=1}^{d} \partial_{x_i} b_i \in C^2(\Omega).$$

A typical example satisfying all assumptions above is the Dirichlet-Laplacian, which we will consider in some our numerical simulations.

We recall that the Hausdorff distance between two compact subsets $K_1, K_2$ of $\mathbf{R}^d$ is defined by

$$d_{\mathcal{H}}(K_1, K_2) = \max\Big(\sup_{x\in K_2} d_{K_1}(x), \sup_{x\in K_1} d_{K_2}(x)\Big).$$

## 12.1.2 Setting

Let $L \in (0, 1)$. We define the set of admissible shapes as

$$\mathcal{U}_L = \{\omega \subset \Omega \text{ measurable } \mid |\omega| \leq L|\Omega|\}. \tag{12.2}$$

**Dynamical optimal shape design problem ($\mathbf{DSD_T}$)**
Let $y_0 \in L^2(\Omega)$, and let $\gamma_1 \geq 0$ and $\gamma_2 \geq 0$ be arbitrary. We consider the parabolic equation controlled by a (measurable) time-varying map $t \mapsto \omega(t)$ of subdomains

$$\partial_t y + Ay = \chi_{\omega(\cdot)}, \quad y_{|\partial\Omega} = 0, \quad y(0) = y_0. \tag{12.3}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Given $T > 0$ and $y_d \in L^2(\Omega)$, we consider the dynamical optimal shape design problem ($\mathbf{DSD_T}$) of determining a measurable path of shapes $t \mapsto \omega(t) \in \mathcal{U}_L$ that minimizes the cost functional

$$J_T(\omega) = \frac{\gamma_1}{2T} \int_0^T \|y(t) - y_d\|^2 \, dt + \frac{\gamma_2}{2} \|y(T) - y_d\|^2,$$

where $y = y(t,x)$ is the solution of (12.3) corresponding to $\omega$.

**Static optimal shape design problem (SSD)**
For the same target function $y_d \in L^2(\Omega)$, we consider the associated static shape design problem

$$\min_{\omega \in \mathcal{U}_L} \frac{\gamma_1}{2} \|y - y_d\|^2, \quad Ay = \chi_\omega, \quad y_{|\partial\Omega} = 0. \tag{SSD}$$

## 12.1.3 Results

### 12.1.3.1 Existence of solutions

The existence and uniqueness of solutions have been established in [12]. To facilitate the reading, we give a sketch of proof in the static case in Section 12.A.3. In few words, we consider a relaxation of ($\mathbf{DSD_T}$) and (**SSD**) with the following idea: we replace the source term $\chi_\omega$, the characteristic function of some shape $\omega \in \mathcal{U}_L$, by a function $a \in L^\infty(\Omega, [0,1])$ (*convexification*, also called *relaxation* in classical optimal shape design theory). Then we can use classical arguments of optimal control theory such as the use of the adjoint variable and the application of first-order optimality conditions coming from the Pontryagin maximum principle (see [16, Chapter 2, Theorem 1.4; Chapter 3, Theorem 2.1]). Convexification methods can be found in [1] and are a particular case of homogenization methods in shape design. which are also very powerful for numerical design. In Section 12.2.1, we introduce the convexified (relaxed) version of problems ($\mathbf{DSD_T}$) and (**SSD**). The existence of optimal shapes is ensured under appropriate assumptions (see [12] for complete proofs).

**Theorem 12.1.1** ([12]). *We distinguish between Lagrange and Mayer cases.*

1. $\gamma_1 = 0$, $\gamma_2 = 1$ *(Mayer case): If $A$ is analytic hypoelliptic in $\Omega$, then there exists a unique optimal shape solution $\omega_T$ of* ($\mathbf{DSD_T}$).
2. $\gamma_1 = 1$, $\gamma_2 = 0$ *(Lagrange case): Assuming that $y_0 \in D(A)$ and that $y_d \in H^2(\Omega)$:*
   (i) *If $y_d < y^0$ or $y_d > y^1$, then there exist unique optimal shape solutions $\bar{\omega}$ and $\omega_T$ of* (**SSD**) *and* ($\mathbf{DSD_T}$), *respectively.*
   (ii) *There exists a function $\beta$ such that if $Ay_d \le \beta$, then there exists a unique optimal shape solution $\bar{\omega}$ of* (**SSD**).

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Some illustrations are provided in Section 12.2.2. Under the assumptions of Theorem 12.1.1, we next derive turnpike properties.

#### 12.1.3.2 Turnpike

The turnpike phenomenon was first observed and investigated by economists for discrete-time optimal control problems (see [7, 17]). There are several possible notions of turnpike properties, some of them being stronger than the others (see [27]). *Exponential turnpike* properties have been established in [9, 19, 20, 23–25] for the optimal triple resulting of the application of Pontryagin's maximum principle, ensuring that the extremal solution (state, adjoint, and control) remains exponentially close to an optimal solution of the corresponding static controlled problem, except at the beginning and the end of the time interval, as soon as $T$ is large enough. This follows from hyperbolicity properties of the Hamiltonian flow. In [12], we establish some results on turnpike property for linear parabolic optimal shape design problem. We denote by $(y_T, p_T, \chi_{\omega_T})$ and $(\bar{y}, \bar{p}, \chi_{\bar{\omega}})$ the optimal triples of both problems ($\mathbf{DSD_T}$) and (**SSD**). We show that the dynamical optimal triple, when $T$ is large, remains most of the time "close" to the static optimal triple. In the Mayer case, we show that the Hausdorff distance between the dynamical optimal shape and the static one satisfies an exponential turnpike.

**Theorem 12.1.2** ([12]). *For $\gamma_1 = 0$, $\gamma_2 = 1$ (Mayer case), $\Omega$ with $C^2$ boundary, and $c = 0$, there exist $T_0 > 0$, $M > 0$, and $\mu > 0$ such that, for every $T \geq T_0$,*

$$d_{\mathcal{H}}(\omega_T(t), \bar{\omega}) \leq Me^{-\mu(T-t)} \quad \forall t \in (0, T).$$

Concerning the Lagrange case, an exponential turnpike property is conjectured, supported by several numerical simulations (see Section 12.3), but for theoretical results, we only establish integral and measure turnpike properties.

**Theorem 12.1.3** ([12]). *For $\gamma_1 = 1$ and $\gamma_2 = 0$ (Lagrange case), there exists $M > 0$ (independent of the final time $T$) such that*

$$\int_0^T \left( \|y_T(t) - \bar{y}\|^2 + \|p_T(t) - \bar{p}\|^2 \right) dt \leq M \quad \forall T > 0.$$

**Theorem 12.1.4** ([12]). *For $\gamma_1 = 1$ and $\gamma_2 = 0$ (Lagrange case), the state-adjoint pair $(y_T, p_T)$ satisfies the* state-adjoint measure-turnpike property*: for every $\varepsilon > 0$, there exists $\Lambda(\varepsilon) > 0$, independent of $T$, such that*

$$|P_{\varepsilon,T}| < \Lambda(\varepsilon) \quad \forall T > 0,$$

*where $P_{\varepsilon,T} = \{t \in [0, T] \mid \|y_T(t) - \bar{y}\| + \|p_T(t) - \bar{p}\| > \varepsilon\}$.*

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Proofs are given in [12], but for the convenience of the reader, we give a sketch of a proof of Theorem 12.1.3 in Appendix 12.A.3. Based on the numerical simulations presented in Section 12.3, we conjecture that the exponential turnpike property is satisfied, i. e., given optimal triples $(y_T, p_T, \chi_{\omega_T})$ and $(\bar{y}, \bar{p}, \bar{\omega})$, there exist $C_1 > 0$ and $C_2 > 0$, independent of $T$, such that

$$\|y_T(t) - \bar{y}\| + \|p_T(t) - \bar{p}\| + \|\chi_{\omega_T(t)} - \chi_{\bar{\omega}}\| \leq C_1(e^{-C_2 t} + e^{-C_2(T-t)})$$

for a. e. $t \in [0, T]$.

## 12.2 Numerical implementation

Computation of the cost function is conditioned to solve the PDE (12.3). We use a variational formulation and a finite element approach. Using **FreeFEM++**, we consider a mesh of the domain $\Omega$, a finite element space with $\mathbb{P}1$-Lagrange elements, and decompose all functions according to this triangulation. To solve numerically ($\mathbf{DSD_T}$) and (**SSD**), we consider the convexification method, which introduces a more general problem, whose solution is expected to be a shape under the assumptions of Theorem 12.1.1 and whose solution is computed via the interior point optimization method (IpOpt, see [26]).

### 12.2.1 Convexification method

The convexification method is a relaxation of the initial problem, which allows us to use classical tools of PDE optimal control. This is a well-known method in shape optimization, which has the double benefit to provide strategies to set theoretical results (see Section 12.A.3 and [21]) and a framework for numerical solving. Given any measurable subset $\omega \subset \Omega$, we identify $\omega$ with its indicator (characteristic) function $\chi_\omega \in L^\infty(\Omega; \{0, 1\})$, and following [2, 21, 22], we identify $\mathcal{U}_L$ with a subset of $L^\infty(\Omega)$. Then the convex closure of $\mathcal{U}_L$ in the $L^\infty$ weak-star topology is the set

$$\overline{\mathcal{U}}_L = \left\{ a \in L^\infty(\Omega; [0, 1]) \;\middle|\; \int_\Omega a(x)\, dx \leq L|\Omega| \right\}, \tag{12.4}$$

which is weak-star compact. More generally, we refer the reader to [1] for details on convexification and homogenization methods. We define the *convexified* (or *relaxed*) optimal control problem $(\mathbf{OCP})_\mathbf{T}$ as the problem of determining a control $t \mapsto a(t) \in \overline{\mathcal{U}}_L$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

minimizing the cost

$$J_T(a) = \frac{\gamma_1}{2T}\int_0^T \|y(t) - y_d\|^2\,dt + \frac{\gamma_2}{2}\|y(T) - y_d\|^2$$

under the dynamical constraints

$$\partial_t y + Ay = a, \quad y_{|\partial\Omega} = 0, \quad y(0) = y_0. \tag{12.5}$$

The corresponding convexified static optimization problem is

$$\min_{a\in\overline{\mathcal{U}}_L} \frac{\gamma_1}{2}\|y - y_d\|^2, \quad Ay = a, \quad y_{|\partial\Omega} = 0. \tag{**SOP**}$$

Solving $(\mathbf{OCP})_\mathbf{T}$ and (**SOP**) under the assumptions of Theorem 12.1.1 implies that the optimal solutions of these convexified problems happen to be the upper level sets of the adjoint state and hence are shapes that are the optimal solutions of $(\mathbf{DSD_T})$ and (**SSD**), respectively. When the assumptions done in Theorem 12.1.1 are not satisfied, a relaxation phenomenon might be observed, namely, it may happen that the optimal solution of $(\mathbf{OCP})_\mathbf{T}$ or (**SOP**) is not the characteristic function of some subset, insofar it takes values in $(0,1)$ on a subset of positive measure. In Section 12.2.2, we give an example of a target function $y_d$ such that the optimal solution $\bar{a}$ of (**SOP**) is not 0 or 1.

Since $(\mathbf{OCP})_\mathbf{T}$ and (**SOP**) are almost linear quadratic, we generate a mesh of $\Omega$ and introduce the finite element space $V_h$ of $\mathbb{P}1$-Lagrange finite elements with its basis $(\omega_i) \in H^1(\Omega)$. Introducing the matrices

$$A_{ij} = \int_\Omega \nabla\omega_i \cdot \nabla\omega_j\,dx, \quad B_{ij} = \int_\Omega \omega_i\omega_j\,dx, \quad L_{v_i} = \int_\Omega \omega_i\,dx,$$

we write (**SOP**) as

$$\min_{(Y,U)\in\mathbf{R}^{2N_v}} (Y - Y_d)^T B(Y - Y_d), \quad AY = BU, \quad L_v^T U \le L|\Omega|, \tag{12.6}$$

where $N_v$ is the size of the finite element space (for $\mathbb{P}1$-Lagrange elements, $N_v$ is the number of vertices). For time discretization, we use an implicit Euler scheme with $N_t$ steps. Introducing the matrices

$$A_t = A + \frac{1}{\delta t}B, \quad B_t = \frac{1}{\delta t}B, \quad \delta t = \frac{T}{N_t - 1}, \tag{12.7}$$

$$\tilde{A} = \begin{pmatrix} I_{N_v} & 0_{N_v} & \cdots & 0_{N_v} \\ B_t & A_t & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0_{N_v} \\ 0_{N_v} & \cdots & B_t & A_t \end{pmatrix}, \quad \tilde{B} = \begin{pmatrix} 0_{N_v} & 0_{N_v} & \cdots & 0_{N_v} \\ 0_{N_v} & B_t & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0_{N_v} \\ 0_{N_v} & \cdots & 0_{N_v} & B_t \end{pmatrix}, \tag{12.8}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

we write $(\mathbf{OCP})_{\mathbf{T}}$ as

$$\min\{(Y - Y_d)^T \tilde{B}(Y - Y_d) \mid (Y, U) \in \mathbf{R}^{2N_v N_t}, \quad \tilde{A}Y = \tilde{B}U, \\ L_v^T U_i \le L|\Omega| \quad \forall i \in \{1, \dots, N_t\}\}. \tag{12.9}$$

Here the state variable $Y$ is seen as an optimization variable. The discretized convexified problems are linear quadratic optimization problems with inequality and equality constraints. Then it is easy to compute the derivatives of the various functions involved. We implement them and their derivatives, and we call the optimization routine IpOpt via **FreeFEM++**. The more information on derivatives and Hessians we give to IpOpt, the faster solving is. To solve (12.6) and (12.9), we compute the gradient and Hessian of the cost function and the Jacobian of the constraint function, and we add (dummy) bound variables to help the solver.

## 12.2.2 Some numerical examples

Here we present some particular cases of the following optimization problem (**SSD**):

$$\min_{\omega} \frac{1}{2}\|y - y_d\|^2, \quad Ay = \chi_\omega, \quad y_{|\partial\Omega} = 0, \quad |\omega| \le L|\Omega|$$

for various operators $A$, spaces $\Omega$, and target functions $y_d$.

### 12.2.2.1 $\Omega = [-1, 1]^2$, $y_d = 0.1$, and $A = -\Delta$

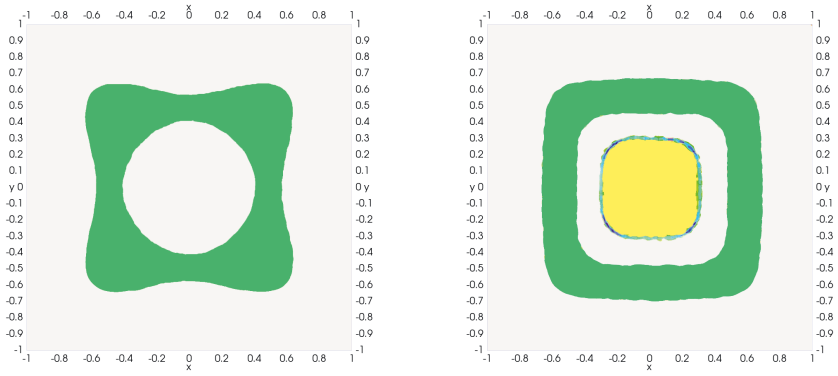
We solve the problem for $y_d = 0.1$ and $y_d = \frac{2-x^2+y^2}{20}$. We plot both solutions in Figure 12.1. The function $y_d$ satisfies the assumptions of Theorem 12.1.1 in the first example and does not in the second one. The convexification method does not require many iterations. This may be due to two facts: first, IpOpt is known to be a very efficient optimization method; second, we have implemented the Hessian of the cost function, and we thus use a Newton method. We notice that in the second case where $y_d = \frac{2-x^2+y^2}{20}$ is no more convex, the optimal solution is not 0 or 1, but takes values in $(0, 1)$.

### 12.2.2.2 $\Omega = [-1, 1]^2$, $y_d = 0.1$, and $A$ is a second-order operator

We focus on the assumptions of Theorem 12.1.1 which are not always satisfied. We take various second-order operators $A$ with ellipticity constants $\theta_1 > 0$, $\theta_1 = 0$, or $\theta_1 < 0$:

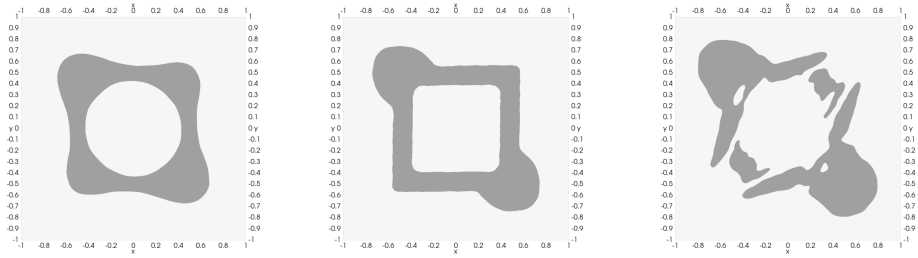
- $\theta_1 > 0$, $Ay = -\partial_{11}y - \partial_{22}y - 0.5\partial_{12}y - 0.5\partial_{21}y$;
- $\theta_1 = 0$, $Ay = -\partial_{11}y - \partial_{22}y - \partial_{12}y - \partial_{21}y$;
- $\theta_1 < 0$, $Ay = -\partial_{11}y - \partial_{22}y - 1.5\partial_{12}y - 1.5\partial_{21}y$,

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use



(a) (b)

**Figure 12.1:** Optimal static shape: (a) $y_d = 0.1$; (b) $y_d = \frac{2-x^2+y^2}{20}$.



(a) (b) (c)

**Figure 12.2:** Optimal static solution: (a) $\theta_1 > 0$; (b) $\theta_1 = 0$; (c) $\theta_1 < 0$.

and plot respective solutions in Figure 12.2. When $\theta_1 > 0$, we are in the hypothesis of Theorem 12.1.1. When $\theta_1 = 0$, it seems that we lose some regularity on the optimal solution. Finally, when $\theta_1 < 0$, we still observe the convergence of our algorithm, but we can no more infer the theoretical results stated at the beginning.

### 12.2.2.3 $\Omega = [-1, 1]^3$, $y_d = 0.025$, and $A = -\Delta$

**FreeFEM**++ is also very powerful to solve PDEs in 3D. Except mesh generation, the previously described methods are unchanged. Of course, in 3D the time required to compute the solution is much larger and will be a limiting factor for solving the time problem. We plot in Figure 12.3 a particular solution of the stationary problem.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

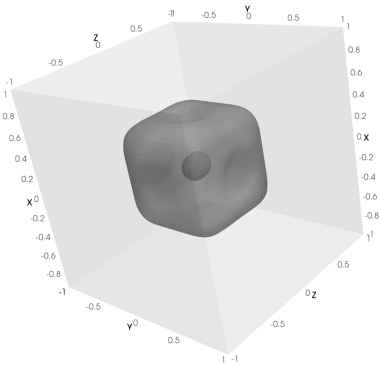


**Figure 12.3:** Optimal static shape 3D.

## 12.3 Numerical examples illustrating the turnpike phenomenon

In this section, to illustrate the turnpike phenomenon, we give numerical simulations for various domains, target functions $y_d$, and operators $A$.

### Case $\Omega = [-1,1]^2$, $y_d = 0.1$, and $A = -\Delta$

We plot on Figure 12.4 the time solution cylinder and the shape behavior at some sample times for several examples. We plot for comparison the time shape at the middle-time of the trajectory, and we observe that it is very similar to the optimal static shape.

To observe the turnpike phenomenon, we plot in Figure 12.5 the error quantity $t \mapsto \|y_T(t) - \bar{y}\| + \|p_T(t) - \bar{p}\| + \|\chi_{\omega_T(t)} - \chi_{\bar{\omega}}\|$ for several final times $T \in \{1, 3, 5\}$.

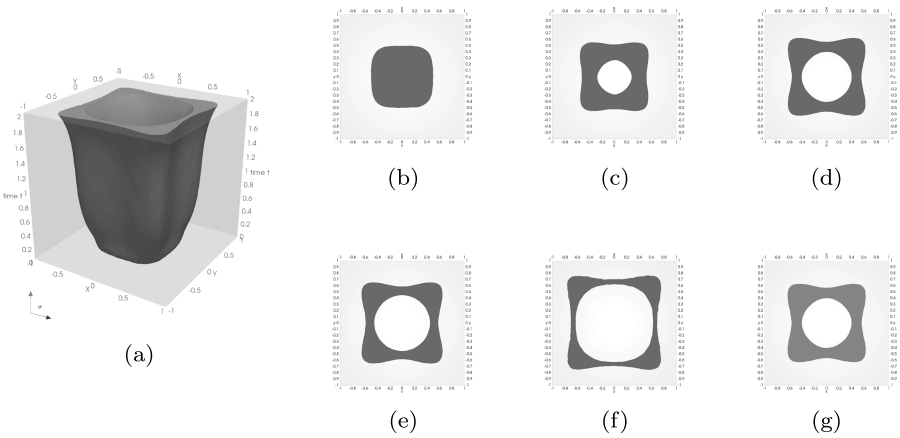



**Figure 12.4:** Optimal dynamical shape: (a) Time shape; (b) $t = 0$; (c) $t = 0.5$; (c) $t \in (0.5, 1.5)$; (e) $t = 1.5$; (f) $t = T$; (g) Static shape.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

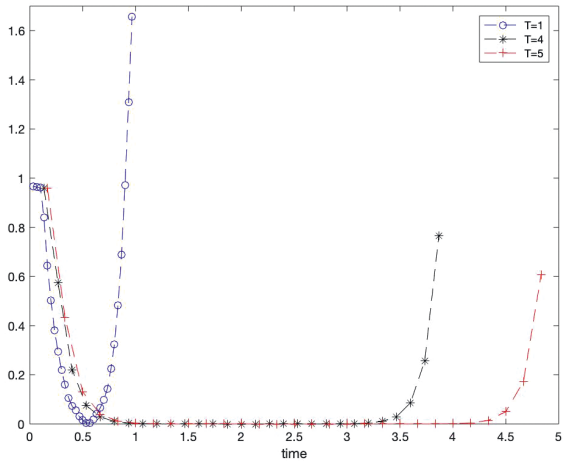



**Figure 12.5:** $t \mapsto \|y_T(t) - \bar{y}\| + \|p_T(t) - \bar{p}\| + \|\chi_{\omega_T(t)} - \chi_{\bar{\omega}}\|$ for $T \in \{1, 3, 5\}$.

The larger is $T$, the more close to 0 is the residual between the two optimal triples. The behavior of the residual function is typical of the turnpike phenomenon and supports our conjecture that exponential turnpike might be proved. Moreover, we plot in Figures 12.6, 12.7, 12.8 and 12.9 solutions for several examples and observe that most of the time the dynamical shape remains very close to the static one. These numerical observations seem to hold systematically under the assumptions of Theorem 12.1.1. The turnpike phenomenon occurs even when we observe relaxation or when the second-order operator $A$ is such that $\theta_1 = 0$.

### Case $\Omega = [-1, 1]^2$, $y_d = \frac{1}{20}(xy + 1)$, and $A = -\Delta$ in Figure 12.6

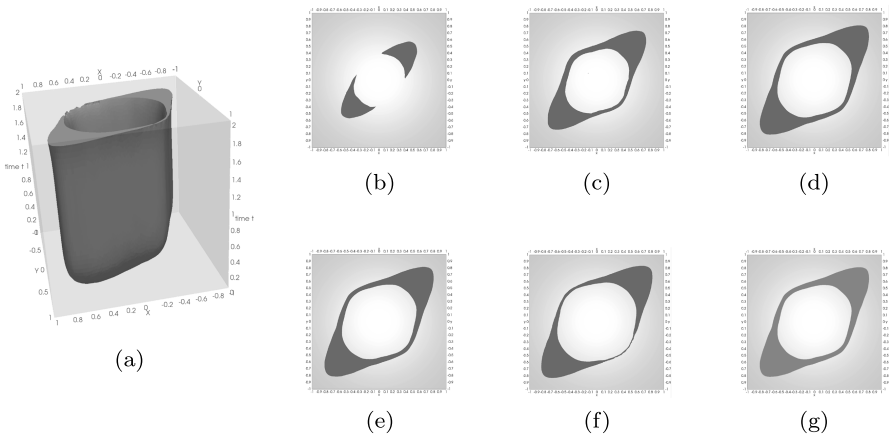



**Figure 12.6:** (a) Time shape; (b) $t = 0$; (c) $t = 0.5$; (c) $t \in (0.5, 1.5)$; (e) $t = 1.5$; (f) $t = T$; (g) Static shape.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Case $\Omega$ = half-stadium, $y_d = \frac{1}{20}(xy + 1)$, and $A = -\Delta$ in Figure 12.7**

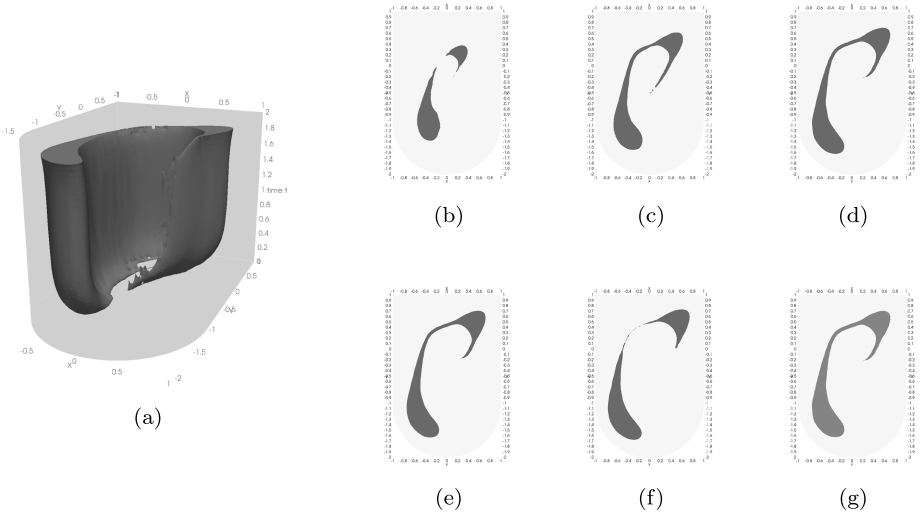



**Figure 12.7:** (a) Time shape; (b) $t = 0$; (c) $t = 0.5$; (c) $t \in (0.5, 1.5)$; (e) $t = 1.5$; (f) $t = T$; (g) Static shape.

**Case $\Omega = [-1, 1]^2$, $y_d = \frac{1}{20}(2\sin(2(x^2 + y^2)) + 1)$, and $Au = -\partial_x((x - y)^2\partial_x u) - \partial_y((x + y)^2\partial_y u)$ in Figure 12.8**

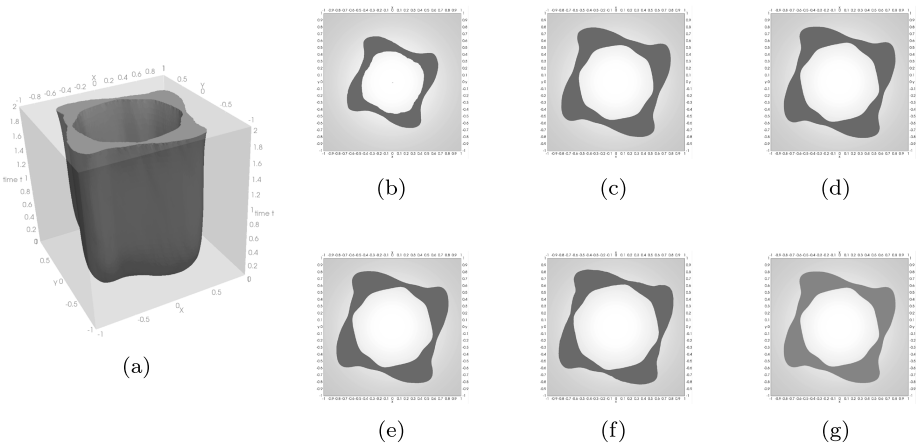



**Figure 12.8:** (a) Time shape; (b) $t = 0$; (c) $t = 0.5$; (c) $t \in (0.5, 1.5)$; (e) $t = 1.5$; (f) $t = T$; (g) Static shape.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Case $\Omega = [-1,1]^3$, $y_d = 0.025$, and $A = -\Delta$ in Figure 12.9**

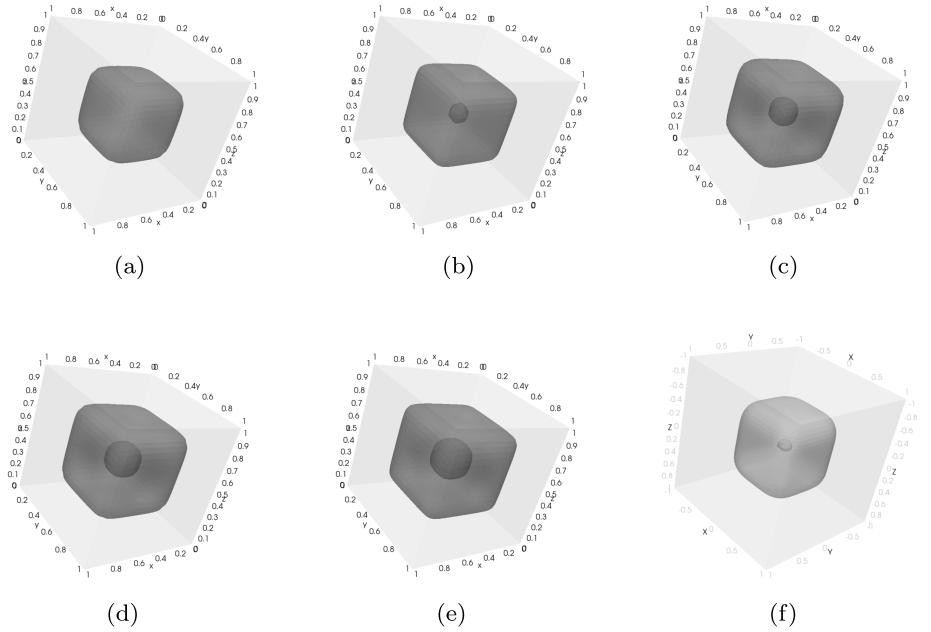



**Figure 12.9:** Time shape: (a) $t = 0$; (b) $t = 0.1$; (b) $t \in ]0.1, 0.9[$; (d) $t = 0.9$; (e) $t = T = 1$; (f) Static shape.

## 12.4 Further comments

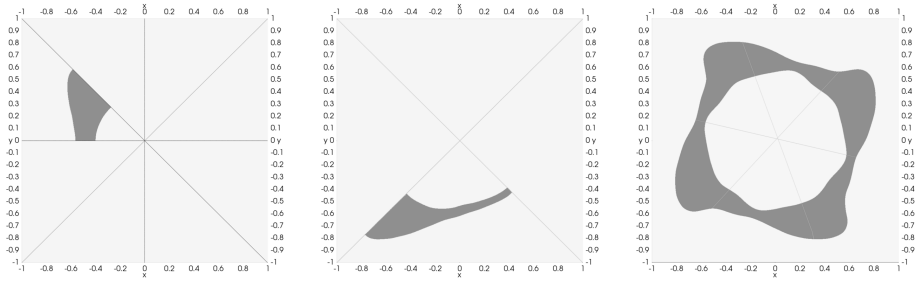
We highlight here some open questions and possible way to pursue the research. Some geometric properties can be conjectured, and we address the question of shape design for semilinear partial differential equations. Finally, we present shortly a shape design problem in the context of shallow water equations.

### 12.4.1 Symmetries of solutions

On the basis of our numerical examples, we conjecture that if $\Omega$, the operator $A$, and the target function $y_d$ share the same symmetry properties, then optimal solutions $\bar{\omega}$ and $\omega_T$ also share symmetry properties.

For instance, Figure 12.10(a) highlights four axes of symmetry ($x = 0$, $y = 0$, $y = x$, $y = -x$), Figure 12.10(b) highlights two axes ($y = x$, $y = -x$), and Figure 12.10(c) shows up a central symmetry property (of center $(0, 0)$ with angle $\pi$).

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use



(a) (b) (c)

**Figure 12.10:** Symmetries of static optimal shapes: (a) $A = -\Delta$, $y_d = 0.1$; (b) $A = -\Delta$, $y_d = \frac{xy+1}{20}$; (c) $Au = -\partial_x((x-y)^2\partial_x u) - \partial_y((x+y)^2\partial_y u)$, $y_d = \frac{1}{20}(2\sin(2(x^2+y^2))+1)$.

## 12.4.2 Semilinear PDEs and numerics

An open and interesting issue is to deal with semilinear PDEs

$$\partial_t y + Ay + f(y) = \chi_\omega, \quad y_{|\partial\Omega} = 0, \quad y(0) = y_0.$$

Theoretical results are not established in this context, but we refer to [10], where the authors are interested in semilinear shape optimization problems, and the existence proofs involve shape optimization problems close to that we deal with in this section. From our side, we provide hereafter several first numerical simulations. We still use the convexification approach combining IpOpt and **FreeFEM**++ for numerical solving. Now, anyway, systems cannot be written as linear quadratic optimization problem as (12.6) and (12.9). We use a fixed point method, so that the solution of the static problem

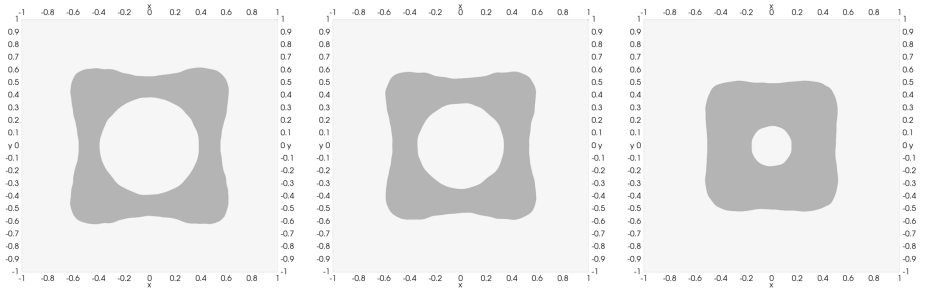
$$Ay + f(y) = \chi_\omega, \quad y_{|\partial\Omega} = 0 \tag{12.10}$$

is sought with an iteration process.

We show in Figure 12.11 some examples of solutions for several functions $f$ and for $A = -\Delta$.

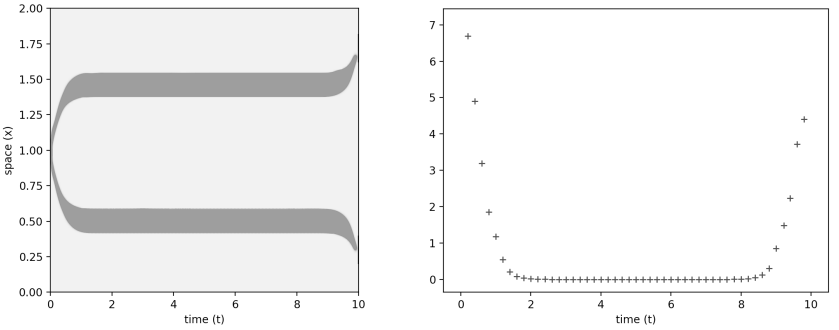
We observe that the optimal solutions are quite similar to those obtained in the linear case. For theoretical results, due to the nonlinearity, new assumptions should be made to ensure the existence and/or uniqueness of solutions. We still observe the turnpike phenomenon for the optimal shape design in 1D. As an example, we take $\Omega = [0,2], A = -\Delta, y_d = 0.1$, and $T = 10$. We plot in Figure 12.12 the dynamical optimal shape, and we observe that it remains most of the time stationary (which is the solution of the corresponding static optimal shape problem). Moreover, we plot the error between both optimal triples, and we still observe a turnpike phenomenon.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use



(a) (b) (c)

**Figure 12.11:** Semilinear optimal designs - stationary case: (a) $f(y) = y(1-y^2)\exp(y)\sin(y)$; (b) $f(y) = (y+0.4)^3$; (c) $f(y) = \frac{1}{y+5}$.




(a) (b)

**Figure 12.12:** Semilinear optimal designs: (a) time-varying optimal shape; (b) error function $t \mapsto \|y_T(t) - \bar{y}\| + \|p_T(t) - \bar{p}\| + \|\chi_{\omega_T(t)} - \chi_{\bar{\omega}}\|$.

## 12.4.3 Optimal design for shallow water equations

### 12.4.3.1 Presentation

We consider the shallow water system in space dimension one, a specific case of shallow water equations, which is a usual way to understand the behavior of a fluid when the height of the water level is much smaller than the other spatial dimensions of the problem, for instance, flows in river or in maritime coast. Using several assumptions (small height of water, hydrostatic pressure, vertical homogeneity of horizontal velocities, etc.), we can derive these from the Euler system. As illustrated in Figure 12.13, we describe the behavior of the fluid at time $t > 0$ and position $x \in \mathbf{R}$ characterized by the water level $h(t,x)$ and the flow $h(t,x)u(t,x)$, and we take into account the variation of the topography $z(t,x)$ and some viscous effects $S_f$. We consider the bottom shape $z$ as the control. For $T > 0$, initial conditions $(h_0, u_0)$, and $(h_d, u_d)$ given as a static wave

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

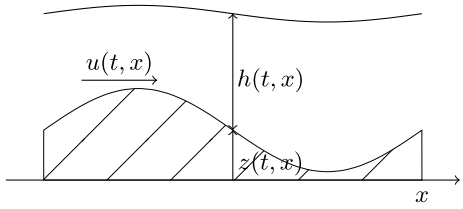



**Figure 12.13:** Shallow water equations.

profile, we search an optimal solution minimizing the functional

$$J_{SW}(z) = \frac{1}{2}\int_0^T \|h(t) - h_d\|^2 + \|u(t) - u_d\|^2 \, dt \tag{12.11}$$

subject to the shallow water equations

$$\begin{aligned} \partial_t h + \partial_x(hu) &= 0, \\ \partial_t(hu) + \partial_x\left(hu^2 + g\frac{h^2}{2}\right) &= -gh\partial_x z + S_f, \\ h(0) = h_0, \quad u(0) &= u_0. \end{aligned} \tag{12.12}$$

Since the target $(h_d, u_d)$ is static, having in mind the phenomenon of static wave occurring in the nature (Eisbach Wave in München or wavemaker in [6]), we would expect that a solution, if it exists, should remain most of the time close to a stationary state solution of a static problem.

### 12.4.3.2 Kinetic interpretation of shallow water equations

To lead this study, we propose a relaxation of the problem via *kinetic equations* (see [4, 5, 18]. We consider a real-valued function $\chi$ of class $C^1$, compactly supported on $\mathbf{R}$, such that

$$\chi(\omega) = \chi(-\omega), \quad \int_{\mathbf{R}} \chi(\omega)\, d\omega = 1, \quad \int_{\mathbf{R}} \omega^2 \chi(\omega)\, d\omega = \frac{g^2}{2}.$$

Defining $c = \sqrt{\frac{gh}{2}}$ and

$$M(t,x,\xi) = \frac{h(t,x)}{c(t,x)}\chi\left(\frac{\xi - u(t,x)}{c(t,x)}\right) = \frac{h(t,x)}{c(t,x)}\chi\left(\frac{\xi - q(t,x)}{h(t,x)c(t,x)}\right), \tag{12.13}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

we get

$$\begin{pmatrix} h \\ q \\ \frac{g}{2}h^2 + \frac{q^2}{h} \end{pmatrix} = \int_{\mathbf{R}} \begin{pmatrix} 1 \\ \xi \\ \xi^2 \end{pmatrix} M(\xi)\, d\xi. \tag{12.14}$$

Following [18], $(h, u)$ is a weak solution of (12.12) if and only if $M$ (defined by (12.13)) satisfies

$$M_t + \xi . M_x - g z_x . M_\xi = Q \tag{12.15}$$

for some $Q$, a collision factor such that $\int_{\mathbf{R}} Q\, d\xi = \int_{\mathbf{R}} \xi Q\, d\xi = 0$.

An open question is to write the initial problem as the minimization of some cost functional

$$\min_z \frac{1}{2} \int_0^T \int_{\mathbf{R}\times\Omega} f^0(x, \xi, M(t, x, \xi))\, d\xi\, dx\, dt$$

with respect to the kinetic equations

$$M_t + \xi . M_x - g z_x . M_\xi = Q, \quad M(0) = \frac{h_0}{c_0} \chi\left( \frac{\xi - u_0}{c_0} \right). \tag{12.16}$$

This would have the double benefit to obtain in an easier way theoretical and numerical results since the nonlinear system (12.12) is replaced by the linear kinetic equation (12.15).

# Appendix

## 12.A.1 Optimality conditions

To show some theoretical results introduced before and in the framework provided by the convexification (see Section 12.2.1), we first give necessary optimality conditions to optimal solutions of the convexified problems stated in [16, Chapters 2 and 3] or [14, Chapter 4] and infer from these necessary conditions that under appropriate assumptions, the optimal controls are indeed characteristic functions.

#### Necessary optimality conditions for $(\mathbf{OCP})_\mathbf{T}$

According to the Pontryagin maximum principle (see [16, Chapter 3, Theorem 2.1]; see also [14]), for any optimal solution $(y_T, a_T)$ of $(\mathbf{OCP})_\mathbf{T}$, one can find an adjoint state

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$p_T \in L^2(0,T;\Omega)$ such that

$$\begin{aligned} &\partial_t y_T + A y_T = a_T, \quad y_{T_{|\partial\Omega}} = 0, \quad y_T(0) = y_0, \\ &\partial_t p_T - A^* p_T = \gamma_1(y_T - y_d), \quad p_{T_{|\partial\Omega}} = 0, \quad p_T(T) = \gamma_2(y_T(T) - y_d), \end{aligned} \tag{12.17}$$

$$\forall a \in \overline{\mathcal{U}}_L, \text{for a. e. } t \in [0,T]: \quad (p_T(t), a_T(t) - a) \geq 0. \tag{12.18}$$

**Necessary optimality conditions for (SOP)**

Similarly, applying [16, Chapter 2, Theorem 1.4], for any optimal solution $(\bar{y}, \bar{a})$ of (**SOP**), there exists an adjoint state $\bar{p} \in L^2(\Omega)$ such that

$$\begin{aligned} A\bar{y} = \bar{a}, \quad \bar{y}_{|\partial\Omega} = 0, \\ -A^*\bar{p} = \gamma_1(\bar{y} - y_d), \quad \bar{p}_{|\partial\Omega} = 0, \end{aligned} \tag{12.19}$$

$$\forall a \in \overline{\mathcal{U}}_L: \quad (\bar{p}, \bar{a} - a) \geq 0. \tag{12.20}$$

Using the bathtub principle (see, e. g., [15, Theorem 1.14]), (12.18) and (12.20) give

$$a_T(\cdot) = \chi_{\{p_T(\cdot) > s_T(\cdot)\}} + c_T(\cdot)\chi_{\{p_T(\cdot) = s_T(\cdot)\}}, \tag{12.21}$$

$$\bar{a} = \chi_{\{\bar{p} > \bar{s}\}} + \bar{c}\chi_{\{\bar{p} = \bar{s}\}} \tag{12.22}$$

with, for a. e. $t \in [0,T]$,

$$c_T(t) \in L^\infty(\Omega;[0,1]) \quad \text{and} \quad \bar{c} \in L^\infty(\Omega;[0,1]), \tag{12.23}$$

$$s_T(\cdot) = \inf\{\sigma \in \mathbf{R} \mid \left|\{p_T(\cdot) > \sigma\}\right| \leq L|\Omega|\}, \tag{12.24}$$

$$\bar{s} = \inf\{\sigma \in \mathbf{R} \mid \left|\{\bar{p} > \sigma\}\right| \leq L|\Omega|\}. \tag{12.25}$$

## 12.A.2 Energy inequalities

We recall some useful inequalities to study the existence and turnpike. Since $\theta$ satisfies (12.1), we can find $\beta > 0$ and $\gamma \geq 0$ such that $\beta \geq \gamma$ and

$$(Au, u) \geq \beta \|u\|^2_{H_0^1(\Omega)} - \gamma \|u\|^2_{L^2(\Omega)}. \tag{12.26}$$

From this there follows the energy inequality (see [8, Chapter 7, Theorem 2]): there exists $C > 0$ such that, for any solution $y$ of (12.5) and almost every $t \in [0,T]$,

$$\left\|y(t)\right\|^2 + \int_0^t \left\|y(s)\right\|^2_{H_0^1(\Omega)} ds \leq C\left(\left\|y_0\right\|^2 + \int_0^t \left\|a(s)\right\|^2 ds\right). \tag{12.27}$$

We improve this inequality, using (12.26) and the Poincaré inequality. We thus find two positive constants $C_1, C_2 > 0$ such that $\frac{d}{dt}\|y(t)\|^2 + C_1\|y(t)\|^2 = f(t) \leq C_2\|a(t)\|^2$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

We solve this differential equation to get $\|y(t)\|^2 = \|y_0\|^2 e^{-C_1 t} + \int_0^t e^{-C_1(t-s)} f(s)\, ds$. Since $f(t) \le C_2 \|a(t)\|^2$ for all $t \in (0,T)$, we obtain that

$$\|y(t)\|^2 \le \|y_0\|^2 e^{-C_1 t} + C_2 \int_0^t e^{-C_1(t-s)} \|a(s)\|^2 \, ds \tag{12.28}$$

for almost every $t \in (0,T)$. The constants $C$, $C_1$, $C_2$ depend only on the domain $\Omega$ (Poincaré inequality) and on the operator $A$ but not on the final time $T$, since (12.26) is satisfied with $\beta \ge \gamma$.

## 12.A.3 Proofs

### 12.A.3.1 Proof of Theorem 12.1.1

Here we first establish the existence of optimal solutions of $(\mathbf{OCP})_\mathbf{T}$ and similarly of (**SOP**). Then we focus on the static case to highlight a key observation to derive the existence of an optimal shape for (**SSD**) and the relaxation phenomenon.

We first see that the infimum exists. Let us take a minimizing sequence

$$(y_n, a_n) \in L^2(0,T; H_0^1(\Omega)) \times L^\infty(0,T; L^2(\Omega, [0,1]))$$

such that, for $n \in \mathbf{N}$ and a. e. $t \in [0,T]$, $a_n(t) \in \overline{\mathcal{U}}_L$, the pair $(y_n, a_n)$ satisfies (12.5), and $J_T(a_n) \to J_T$. The sequence $(a_n)$ is bounded in $L^\infty(0,T; L^2(\Omega,[0,1]))$, so using (12.27) and (12.28), the sequence $(y_n)$ is bounded in $L^\infty(0,T; L^2(\Omega)) \cap L^2(0,T; H_0^1(\Omega))$. We show then, using (12.5), that the sequence $(\frac{\partial y_n}{\partial t})$ is bounded in $L^2(0,T; H^{-1}(\Omega))$. We subtract a sequence, still denoted by $(y_n, a_n)$, such that we can find a pair $(y,a)$ lying in space $L^2(0,T; H_0^1(\Omega)) \times L^\infty(0,T; L^2(\Omega,[0,1]))$ with

$$\begin{aligned} y_n &\rightharpoonup y \quad \text{weakly in } L^2(0,T; H_0^1(\Omega)), \\ \partial_t y_n &\rightharpoonup \partial_t y \quad \text{weakly in } L^2(0,T; H^{-1}(\Omega)), \\ a_n &\rightharpoonup a \quad \text{weakly } \star \text{ in } L^\infty(0,T; L^2(\Omega,[0,1])). \end{aligned} \tag{12.29}$$

We deduce that

$$\begin{aligned} \partial_t y_n + A y_n - a_n \to \partial_t y + A y - a \quad \text{in } \mathcal{D}'((0,T)\times\Omega), \\ y_n(0) \rightharpoonup y(0) \quad \text{weakly in } L^2(\Omega). \end{aligned} \tag{12.30}$$

Using (12.30), we get that $(y,a)$ is a weak solution of (12.5). Moreover, since

$$L^\infty(0,T; L^2(\Omega,[0,1])) = \left(L^1(0,T; L^2(\Omega,[0,1]))\right)'$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

(see [11, Corollary 1.3.22]), convergence (12.29) implies that for every $v \in L^1(0,T)$ satisfying $v \geq 0$ and $\|v\|_{L^1(0,T)} = 1$, we have

$$\int_0^T \left( \int_\Omega a(t,x)\,dx \right) v(t)\,dt \leq L|\Omega|.$$

Since the function $f_a$ defined by

$$f_a(t) = \int_\Omega a(t,x)\,dx$$

belongs to $L^\infty(0,T)$, we have

$$\|f_a\|_{L^\infty(0,T)} = \sup\left\{ \int_0^T \left( \int_\Omega a(t,x)\,dx \right) v(t)\,dt \;\middle|\; v \in L^1(0,T),\ \|v\|_{L^1(0,T)} = 1 \right\}.$$

Therefore $\|f_a\|_{L^\infty(0,T)} \leq L|\Omega|$ and $\int_\Omega a(t,x)\,dx \leq L|\Omega|$ for a. e. $t \in (0,T)$. This shows that the pair $(y,a)$ is admissible. Since $H_0^1(\Omega)$ is compactly embedded in $L^2(\Omega)$, using the Aubin–Lions compactness lemma (see [3]), we obtain

$$y_n \to y \quad \text{strongly in } L^2(0,T;L^2(\Omega)).$$

Then by the weak lower semicontinuity of $J_T$ and by the Fatou lemma we get that

$$J_T(a) \leq \liminf J_T(a_n).$$

Hence $a$ is an optimal control for $(\mathbf{OCP})_\mathbf{T}$, which we rather denote by $a_T$ (and $\bar{a}$ for (**SOP**)).

Let us now focus on the part 2(ii) of Theorem 12.1.1. We focus on $\gamma_1 = 1$, $\gamma_2 = 0$ (Lagrange case), and (**SSD**). Since $\bar{a}$ is a s solution of (**SOP**), it satisfies the optimality conditions stated in (12.19)–(12.22). One key observation noting that if $|\{\bar{p} = \bar{s}\}| = 0$, then from (12.22) it follows that the static optimal control $\bar{a}$ is in fact the characteristic function of a shape $\bar{\omega} \in \mathcal{U}_L$. With this in mind, we give a useful lemma.

**Lemma 12.A.1** ([13, Theorem 3.2]). *Given any $p \in [1,+\infty)$ and any $u \in W^{1,p}(\Omega)$ such that $|\{u = 0\}| > 0$, we have $\nabla u = 0$ a. e. on $\{u = 0\}$.*

*2(ii)* We assume that $Ay_d \leq \beta$ in $\Omega$ with $\beta = \bar{s}Ac^*$. Having in mind (12.19) and (12.22), we assume by contradiction that $|\{\bar{p} = \bar{s}\}| > 0$. Since $A$ and $A^*$ are differential operators, applying $A^*$ to $\bar{p}$ on $\{\bar{p} = \bar{s}\}$, we obtain by Lemma 12.A.1 that

$$A^*\bar{p} = c^*\bar{s} \quad \text{on } \{\bar{p} = \bar{s}\}.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Since $(\bar{y}, \bar{p})$ satisfies (12.19), we get

$$y_d - \bar{y} = c^* \bar{s} \quad \text{on } \{\bar{p} = \bar{s}\}.$$

Then applying $A$ to this equation, we get that

$$Ay_d - \bar{s}Ac^* = A\bar{y} = \bar{a} \quad \text{on } \{\bar{p} = \bar{s}\}.$$

Therefore

$$Ay_d - \bar{s}Ac^* \in (0,1) \quad \text{on } \{\bar{p} = \bar{s}\},$$

which contradicts $Ay_d \leq \beta$. Hence $|\{\bar{p} = \bar{s}\}| = 0$, and thus (12.22) implies $\bar{a} = \chi_{\bar{\omega}}$ for some $\bar{\omega} \in \mathcal{U}_L$. The existence of solution for (**SSD**) is proved.

The uniqueness of optimal controls comes from the strict convexity of the cost functionals. Indeed, in the dynamical case, whatever $(\gamma_1, \gamma_2) \neq (0,0)$ may be, $J_T$ is strictly convex with respect to variable $y$. The injectivity of the control-to-state mapping gives the strict convexity with respect to the variable $a$. In addition, the uniqueness of $(\bar{y}, \bar{p})$ follows by application of the Poincaré inequality, and the uniqueness of $(y_T, p_T)$ follows from the Grönwall inequality (12.28) in Appendix 12.A.2.

**Remark 12.A.2.** Condition 2(ii) in Theorem 12.1.1 is a necessary condition. We can construct an example where $Ay_d \leq \beta$ is not satisfied and where we observe relaxation, which is closely related to the fact $|\{\bar{p} = \bar{s}\}| > 0$.

Indeed, we plot on Figure 12.14 the adjoint state $\bar{p}$ for the static problem in 1D. At the left-hand side, $\bar{p}$ is assumed to be analytic: in this case, all level sets of $\bar{p}$ have zero Lebesgue measure (there is no subset of positive measure on which $\bar{p}$ would remain constant). When $\bar{p}$ is not analytic and remains constant on a subset of positive measure (see Figure 12.14 in red), we do not have necessarily zero Lebesgue measure level sets, and on $\{\bar{p} = \bar{s}\}$, $\bar{a}$ can take values in $(0,1)$.

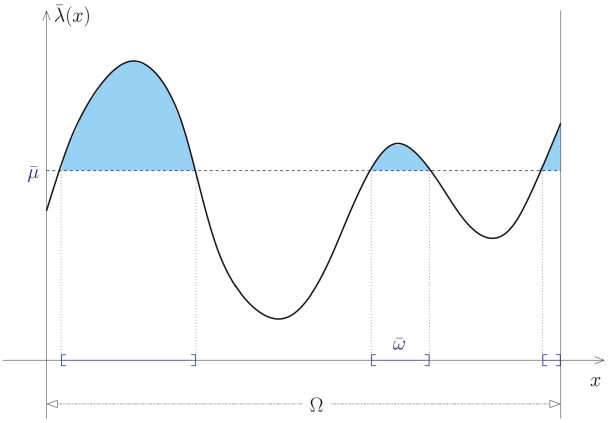
*Proof of Theorem* 12.1.3. For $\gamma_1 = 1$ and $\gamma_2 = 0$ (Lagrange case), the cost is

$$J_T(\omega) = \frac{1}{2T} \int_0^T \left\| y(t) - y_d \right\|^2 dt.$$

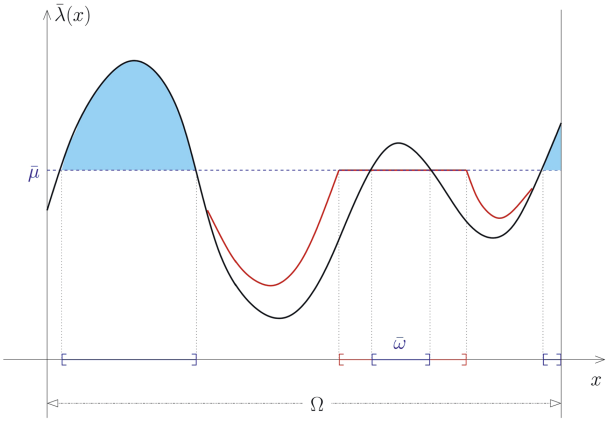
We consider the triples $(y_T, p_T, \chi_{\omega_T})$ and $(\bar{y}, \bar{p}, \chi_{\bar{\omega}})$ satisfying the optimality conditions (12.17) and (12.19). Since $\chi_{\omega_T(t)}$ is bounded at each time $t \in [0,T]$, by an application of the Grönwall inequality (12.28) in Appendix 12.A.2 to $y_T$ and $p_T$ we can find a constant $C > 0$, depending only on $A$, $y_0$, $y_d$, $\Omega$, $L$, such that

$$\forall T > 0, \quad \left\| y_T(T) \right\|^2 \leq C \quad \text{and} \quad \left\| p_T(0) \right\|^2 \leq C.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use




(a)




(b)

**Figure 12.14:** Optimal shape design existence and relaxation: (a) No relaxation: Shape existence; (b) Relaxation.

Setting $\tilde{y} = y_T - \bar{y}, \tilde{p} = p_T - \bar{p}, \tilde{a} = \chi_{\omega_T} - \chi_{\bar{\omega}}$, we have

$$\partial_t \tilde{y} + A\tilde{y} = \tilde{a}, \quad \tilde{y}_{|\partial\Omega} = 0, \quad \tilde{y}(0) = y_0 - \bar{y}, \tag{12.31}$$

$$\partial_t \tilde{p} - A^* \tilde{p} = \tilde{y}, \quad \tilde{p}_{|\partial\Omega} = 0, \quad \tilde{p}(T) = -\bar{p}. \tag{12.32}$$

First, using (12.17) and (12.19), we have $(\tilde{p}(t), \tilde{a}(t)) \geq 0$ for almost every $t \in [0, T]$. Multiplying (12.31) by $\tilde{p}$, (12.32) by $\tilde{y}$, and then adding them, we can use the fact that

$$(\bar{y} - y_0, \tilde{p}(0)) - (\tilde{y}(T), \bar{p}) = \int_0^T (\tilde{p}(t), \tilde{a}(t))\, dt + \int_0^T \|\tilde{y}(t)\|^2\, dt.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

By the Cauchy–Schwarz inequality we get a new constant $C > 0$ such that

$$\frac{1}{T}\int_0^T \|\tilde{y}(t)\|^2\,dt + \frac{1}{T}\int_0^T (\tilde{p}(t), \tilde{a}(t))\,dt \le \frac{C}{T}.$$

The two terms at the left-hand side are positive, and using inequality (12.27) with the function $\zeta(t) = \tilde{p}(T-t)$, we finally obtain $M > 0$, independent of $T$, such that

$$\frac{1}{T}\int_0^T \left(\|y_T(t) - \bar{y}\|^2 + \|p_T(t) - \bar{p}\|^2\right) dt \le \frac{M}{T}.$$

□

# Bibliography


[1] G. Allaire. *Conception Optimale de Structures*, volume 58 of Mathématiques & Applications (Berlin) [Mathematics & Applications]. Springer-Verlag, Berlin, 2007. With the collaboration of Marc Schoenauer (INRIA) in writing Chapter 8.

[2] G. Allaire, A. Münch, and F. Periago. Long time behavior of a two-phase optimal design for the heat equation. *SIAM J. Control Optim.*, 48(8):5333–5356, 2010.

[3] J. Aubin. Un théorème de compacité. *C. R. Acad. Sci. Paris*, 256:5042–5044, 1963.

[4] E. Audusse, F. Bouchut, M.-O. Bristeau, and J. Sainte-Marie. Kinetic entropy inequality and hydrostatic reconstruction scheme for the Saint-Venant system. *Math. Comput.*, 85(302):2815–2837, 2016.

[5] M.-O. Bristeau and B. Coussin. Boundary conditions for the shallow water equations solved by kinetic schemes. Research Report RR-4282, INRIA, 2001. Projet M3N.

[6] Citywave. https://citywave.de/fr/.

[7] R. Dorfman, P. A. Samuelson, and R. M. Solow. *Linear Programming and Economic Analysis*. A Rand Corporation Research Study. McGraw-Hill Book Co., Inc., New York–Toronto–London, 1958.

[8] L. C. Evans. *Partial Differential Equations*, 2nd edition, volume 19 of Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2010.

[9] L. Grüne, M. Schaller, and A. Schiela. Exponential sensitivity and turnpike analysis for linear quadratic optimal control of general evolution equations, Dezember 2018.

[10] A. Henrot, I. Mazari, and Y. Privat. Shape optimization of a Dirichlet type energy for semilinear elliptic partial differential equations. Working paper or preprint, Apr. 2020.

[11] T. Hytönen, J. van Neerven, M. Veraar, and L. Weis. *Analysis in Banach spaces. Vol. I. Martingales and Littlewood–Paley Theory*, volume 63 of Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]. Springer, Cham, 2016.

[12] G. Lance, E. Trélat, and E. Zuazua. Shape turnpike for linear parabolic PDE models. *Syst. Control Lett.*, 142:104733, 9, 2020.

[13] H. Le Dret. *Nonlinear Elliptic Partial Differential Equations: An introduction*. Universitext. Springer, Cham, 2018. Translated from the 2013 French edition [MR3235838].

[14] X. J. Li and J. M. Yong. *Optimal Control Theory for Infinite-Dimensional Systems*. Systems & Control: Foundations & Applications. Birkhäuser Boston, Inc., Boston, MA, 1995.


EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[15] E. H. Lieb and M. Loss. *Analysis*, 2nd edition, volume 14 of Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2001.
[16] J.-L. Lions. *Optimal Control of Systems Governed by Partial Differential Equations*, volume 170 of Die Grundlehren der mathematischen Wissenschaften. Springer-Verlag, New York–Berlin, 1971. Translated from the French by S. K. Mitter.
[17] L. W. McKenzie. Turnpike theorems for a generalized Leontief model. *Econometrica*, 31(1/2):165–180, 1963.
[18] B. Perthame and C. Simeoni. A kinetic scheme for the Saint-Venant system with a source term. *Calcolo*, 38(4):201–231, 2001.
[19] A. Porretta and E. Zuazua. Long time versus steady state optimal control. *SIAM J. Control Optim.*, 51(6):4242–4273, 2013.
[20] A. Porretta and E. Zuazua. Remarks on long time versus steady state optimal control. In *Mathematical Paradigms of Climate Science*, volume 15 of Springer INdAM Series, pages 67–89. Springer, [Cham], 2016.
[21] Y. Privat, E. Trélat, and E. Zuazua. Optimal shape and location of sensors for parabolic equations with random initial data. *Arch. Ration. Mech. Anal.*, 216(3):921–981, 2015.
[22] Y. Privat, E. Trélat, and E. Zuazua. Optimal observability of the multi-dimensional wave and Schrödinger equations in quantum ergodic domains. *J. Eur. Math. Soc.*, 18(5):1043–1111, 2016.
[23] E. Trélat and C. Zhang. Integral and measure-turnpike properties for infinite-dimensional optimal control systems. *Math. Control Signals Syst.*, 30(1):Art. 3, 34, 2018.
[24] E. Trélat, C. Zhang, and E. Zuazua. Steady-state and periodic exponential turnpike property for optimal control problems in Hilbert spaces. *SIAM J. Control Optim.*, 56(2):1222–1252, 2018.
[25] E. Trélat and E. Zuazua. The turnpike property in finite-dimensional nonlinear optimal control. *J. Differ. Equ.*, 258(1):81–114, 2015.
[26] A. Wächter and L.-T. Biegler. On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. *Math. Program.*, 106(1):25–57, 2006.
[27] A. J. Zaslavski. *Turnpike Theory of Continuous-Time Linear Optimal Control Problems*, volume 104 of Springer Optimization and Its Applications. Springer, Cham, 2015.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Gabriela Marinoschi

# 13 Feedback stabilization of Cahn–Hilliard phase-field systems

**Abstract:** We provide a compact presentation, relying on the papers [4, 17, 18], of the internal feedback stabilization of phase-field systems of Cahn–Hilliard type by using a feedback controller with support in a subset of the flow domain. Here the feedback stabilization technique is based on the design of the controller as a linear combination of the unstable modes of the corresponding linearized system, followed by its representation in a feedback form by means of an optimization method. Results are provided both for a regular potential involved in the phase field equation (the double-well potential) and for a singular case represented by a logarithmic-type potential. The feedback stabilization is studied in the case with viscosity effects and in the limit case as the viscosity tends to zero.

**Keywords:** feedback control, closed-loop system, stabilization, Cahn–Hilliard system, viscosity effects, logarithmic potential

**MSC 2010:** 93D15, 35K52, 35Q79, 35Q93, 93C20

## 13.1 Introduction

The celebrated Cahn–Hilliard (CH) equation was proposed by Cahn and Hilliard [11] to model the process of the instantaneous separation of a binary mixture in its components. This equation plays an essential role in material science and also in a wide variety of chemical, biological, physical, and engineering fields such as multiphase fluid flows, microstructures with elastic nonhomogeneity, tumor growth simulation, diblock copolymer, spinodal decomposition, image inpainting, and topology optimization (see, e. g., [15] and the references therein).

We deal with the CH system, which describes the evolution of the phase field $\varphi$ and a chemical potential $\mu$ by equations (13.2)–(13.3) below under the Caginalp approach, meaning that it is coupled with equation (13.1) for the temperature dynamics $\theta$ (see [10]),

$$(\theta + l\varphi)_t - \Delta\theta = 0 \quad \text{in } (0,\infty)\times\Omega, \tag{13.1}$$

$$\varphi_t - \Delta\mu = 0 \quad \text{in } (0,\infty)\times\Omega, \tag{13.2}$$

**Gabriela Marinoschi,** Gheorghe Mihoc-Caius Iacob Institute of Mathematical Statistics and Applied Mathematics of the Romanian Academy, Calea 13 Septembrie 13, Bucharest, Romania, e-mail: gabriela.marinoschi@acad.ro

https://doi.org/10.1515/9783110695984-013

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$$\mu = \tau\varphi_t - \nu\Delta\varphi + F'(\varphi) - \gamma\theta \quad \text{in } (0,\infty)\times\Omega. \tag{13.3}$$

Phase separation is the creation of two distinct phases from a single homogeneous mixture. The phase field is assumed to possibly take two distinct values (for instance, +1 and –1) corresponding to the different phases, with a smooth change between these values in a zone around the interface. These equations are completed with initial data and standard homogeneous Neumann boundary conditions

$$\theta(0) = \theta_0, \quad \varphi(0) = \varphi_0 \quad \text{in } \Omega, \tag{13.4}$$

$$\frac{\partial\theta}{\partial n} = \frac{\partial\varphi}{\partial n} = \frac{\partial\mu}{\partial n} = 0 \quad \text{on } (0,\infty)\times\partial\Omega. \tag{13.5}$$

The space domain $\Omega$ is considered a sufficiently regular open bounded connected subset of $\mathbb{R}^d$, $d = 1, 2, 3$, the time $t \in (0,\infty)$, $n$ is the outward normal vector to the boundary, and $l$, $\gamma$, and $\nu$ are positive constants with some physical meaning. Equation (13.3) describes the process in which possible effects of the viscosity of the mixture are taken into account. The mixture viscosity is indicated by the constant $\tau \in [0,1]$. The limit case $\tau = 0$ characterizes the nonviscous Cahn–Hilliard model and expresses a degeneracy in equation (13.3).

As in all models of phase fields, we meet the function denoted here $F'$ that is the derivative of a potential coming from the physical models deduced by the Ginzburg–Landau theory. Standard functions for the potential $F$ are polynomials of even degree with a strictly positive leading coefficient, as, e. g., the double-well potential

$$F(r) = \frac{(r^2-1)^2}{4}, \tag{13.6}$$

the logarithmic potential

$$F(r) = (1+r)\ln(1+r) + (1-r)\ln(1-r) - ar^2 \quad \text{for } r \in (-1,1), \tag{13.7}$$

where $a$ is positive and large enough to prevent the convexity, and subdifferentials of convex lower semicontinuous functions (see, e. g., explanations in [12]).

Equations and conditions (13.1)–(13.5) give rise to the so-called conserved phase-field system, because they lead to the mass conservation of $\varphi$, which is obtained by integrating (13.2) in space and time and using the boundary condition for $\mu$ in (13.5) and the initial condition for $\varphi$ in (13.4).

We investigate the stabilization around a stationary solution $(\theta_\infty, \varphi_\infty)$ to (13.1)–(13.5) and involve two controllers with the support in an open subset $\omega$ of $\Omega$, acting on the right-hand sides of equations (13.1)–(13.2). The aim is to exponentially stabilize this system around $(\varphi_\infty, \theta_\infty)$ using controllers computed in a feedback form, namely to prove that $\lim_{t\to\infty}(\varphi(t), \theta(t)) = (\varphi_\infty, \theta_\infty)$ with exponential convergence rate when the initial datum $(\varphi_0, \theta_0)$ is in a certain neighborhood of $(\varphi_\infty, \theta_\infty)$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

In this survey, based on the papers [4, 17, 18], we discuss separately the previously specified cases of a viscous (nondegenerate) and that of a nonviscous (degenerate) system and reveal in the proofs the particularities related to the double-well and logarithmic potentials.

The proofs consist in a sequence of intermediate results referring to:

- the well-posedness and stabilization of the linearized system by a finite-dimensional control in Propositions 13.2.1 and 13.2.2;
- the representation of the feedback controller by an optimization method and the characterization of its properties in Propositions 13.2.3 and 13.2.4;
- the proof of the existence of a unique solution to the *nonlinear closed-loop system* (with the feedback controller expressed in terms of the solution) and the stabilization of this solution in Theorem 13.2.5.

In Section 13.2, we give these results for the viscous case and a double-well potential. The corresponding facts for the viscous case with logarithmic potential and for the nonviscous case are given in Sections 13.3 and 13.4, respectively, without detailing the proofs, specifying only the differences with respect to the first case.

The technique followed here is the Riccati-based stabilization introduced in [21] and [5] and used and developed later on in [3–8, 20] for Navier–Stokes equations and nonlinear parabolic systems.

We mention that in a surprising way the viscous case with $\tau > 0$ studied in [18], which is associated to a nondegenerate PDE system, is more challenging than the degenerate nonviscous case. Some transformations of the original system lead, in the first case, to an integro-differential system. The linearized system has no longer a self-adjoint operator as in the degenerate case ($\tau = 0$) studied in [4], so that the simplification induced by such an operator cannot be used. Here complex eigenvalues and eigenvectors and, consequently, the stabilization in complexified spaces should be taken into consideration. Results are provided in the three-dimensional case for the system with regular potential (13.6) and in the one-dimensional case for singular potential (13.7). This last part provides a result of feedback stabilization for a regularization $F_\varepsilon$ of the singular potential in Theorem 13.3.1. This stabilization theorem proved for the nonlinear system corresponding to $F_\varepsilon$ can be viewed as a stand-alone result, which can be applied to other models, as for example, for reaction–diffusion processes with nonlinear sources. The stabilization of the system with singular function $F$ is possible and follows on the basis of a compactness result working in one dimension in Theorem 13.3.2.

In the nonviscous case an appropriate function transformation leads to a system with a self-adjoint operator, which simplifies the analysis of the stabilization of the linearized system. The proofs are developed now in real spaces. The representation of the feedback controller and the proof of the stabilization of the nonlinear system are done by similar methods as in the first case. However, in this situation the final result follows under some conditions limiting the magnitude of the gradient and Laplacian

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

of the stationary solution aimed to be stabilized. These are the essential differences between these two cases.

### 13.1.1 Functional framework

Let us consider the standard space triplet $H = L^2(\Omega) \subset V = H^1(\Omega) \subset V' = (H^1(\Omega))'$ and introduce the linear operator $A : D(A) \subset H \to H$,

$$A = I - \tau\Delta \quad \text{with } D(A) = \left\{ w \in H^2(\Omega);\ \frac{\partial w}{\partial n} = 0 \text{ on } \partial\Omega \right\}, \tag{13.8}$$

where $I$ is the identity operator. This operator will be used in the viscous case, $\tau > 0$, and is replaced by $I - \Delta$ in the nonviscous case. The operator $A$ is linear continuous, self-adjoint, and $m$-accretive on $H$, so that its fractional powers $A^\alpha$, $\alpha \geq 0$, can be defined (see, e. g., [19], p. 72). The domain is $D(A^\alpha) = \{w \in H; \|A^\alpha w\|_H < \infty\}$, and the norm $\|w\|_{D(A^\alpha)} = \|A^\alpha w\|_H$. Moreover, $D(A^\alpha) \subset H^{2\alpha}(\Omega)$, with equality if and only if $2\alpha < 3/2$ (see [13]). In the proofs, we will rely on the following inequalities involving the powers of $A$:

$$\|A^\alpha w\|_H \leq C \|A^{\alpha_1} w\|_H^\lambda \|A^{\alpha_2} w\|_H^{1-\lambda} \quad \text{for } \alpha = \lambda\alpha_1 + (1-\lambda)\alpha_2,\ \lambda \in [0,1], \tag{13.9}$$

$$\|A^\alpha w\|_H \leq C \|A^\beta w\|_H \quad \text{if } \alpha < \beta, \tag{13.10}$$

$$\|A^\alpha w\|_{H^\beta(\Omega)}^2 \leq C \|A^{\alpha+\beta/2} w\|_H^2, \tag{13.11}$$

with $C$ depending on the domain $\Omega$ and the exponents.

## 13.2 The viscous case with the double-well potential

In (13.1)–(13.5) where $F$ is given by (13.6) we make the function transformation $\theta = \sigma - l\varphi$ and replace the expression of $\mu$ into (13.1). The controlled system we will study is

$$(1 - \tau\Delta)\varphi_t + \nu\Delta^2\varphi - \Delta F'(\varphi) - \gamma l\Delta\varphi + \gamma\Delta\sigma = (1 - \tau\Delta)(f_\omega v) \quad \text{in } (0,\infty)\times\Omega, \tag{13.12}$$

$$\sigma_t - \Delta\sigma + l\Delta\varphi = f_\omega u \quad \text{in } (0,\infty)\times\Omega, \tag{13.13}$$

$$\varphi(0) = \varphi_0, \quad \sigma(0) = \sigma_0 := \theta_0 + l\varphi_0 \quad \text{in } \Omega, \tag{13.14}$$

$$\frac{\partial\varphi}{\partial n} = \frac{\partial\Delta\varphi}{\partial n} = \frac{\partial\sigma}{\partial n} = 0 \quad \text{in } (0,\infty)\times\partial\Omega, \tag{13.15}$$

where the second boundary condition in (13.15) follows by (13.5). The function $f_\omega$ is taken such that

$$f_\omega \in C_0^\infty(\Omega), \quad \text{supp} f_\omega \subset \omega, \quad f_\omega > 0 \quad \text{on } \omega_0, \tag{13.16}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

where $\omega_0$ is an open subset of positive measure of $\omega$. The particular form of the control $(1 - \tau\Delta)(f_\omega v)$ was set to ensure that this controller has the support in $\omega_0$, and also for technical considerations.

The feedback stabilization of the Cahn–Hilliard system in the viscous case for both regular potential (13.6) and singular potential (13.7) was treated in [18].

Although the proof of the existence of a solution to the stationary uncontrolled system (13.12)–(13.15) is beyond the aim of this work, we state that such a solution exists and has the properties: $\theta_\infty$ is constant, and $\varphi_\infty \in H^4(\Omega) \subset C^2(\overline{\Omega})$ if $F$ is the double-well potential (see [4], Lemma A1 in Appendix). For the logarithmic case, it is sufficient to observe that always there are constant stationary solutions $\theta_\infty$ and $\varphi_\infty$.

As usual, we will study the stabilization of the null solution to the system below written for the functions $y := \varphi - \varphi_\infty$, $z := \sigma - \sigma_\infty$:

$$(1 - \tau\Delta)y_t + \nu\Delta^2 y - \Delta(F'(y + \varphi_\infty) - F'(\varphi_\infty)) - \gamma l\Delta y + \gamma\Delta z = (1 - \tau\Delta)(f_\omega v), \tag{13.17}$$

$$z_t - \Delta z + l\Delta y = f_\omega u \quad \text{in } (0,\infty)\times\Omega, \tag{13.18}$$

$$y(0) = y_0 = \varphi_0 - \varphi_\infty, \quad z(0) = z_0 = \sigma_0 - \sigma_\infty, \tag{13.19}$$

$$\frac{\partial y}{\partial n} = \frac{\partial \Delta y}{\partial n} = \frac{\partial z}{\partial n} = 0. \tag{13.20}$$

Applying in (13.17) the Taylor expansion for $F'(y + \varphi_\infty)$ around $\varphi_\infty$, we obtain

$$(1 - \tau\Delta)y_t + \nu\Delta^2 y - \Delta(F''(\varphi_\infty)y) - \gamma l\Delta y + \gamma\Delta z = (1 - \tau\Delta)(f_\omega v) + \Delta F_r(y), \tag{13.21}$$

where the nonlinear rest of second order of the Taylor expansion, $F_r(y)$, was moved on the right-hand side. This expansion can be written for the functions considered before (polynomial and logarithmic) under certain hypotheses in the second case (these will be specified in Section 13.3).

We write (13.21) and (13.18)–(13.20) in terms of $A$ (by replacing $\Delta = \frac{1}{\tau}(I - A)$), and since $A$ is surjective, we apply $A^{-1}$ in (13.21). The stabilization of system (13.17)–(13.20) is thus reduced to the stabilization of the equivalent integro-differential system

$$\begin{aligned} &y_t + \frac{\nu}{\tau^2}(A + A^{-1} - 2)y - \frac{1}{\tau}(A^{-1} - I)(F''(\varphi_\infty)y) \\ &\quad + \frac{\gamma}{\tau}(A^{-1} - I)z - \frac{\gamma l}{\tau}(A^{-1} - I)y = f_\omega v + \frac{1}{\tau}(A^{-1} - I)F_r(y) \quad \text{in } (0,\infty)\times\Omega, \end{aligned} \tag{13.22}$$

$$z_t + \frac{1}{\tau}(A - I)z + \frac{l}{\tau}(I - A)y = f_\omega u \quad \text{in } (0,\infty)\times\Omega, \tag{13.23}$$

$$y(0) = y_0, \quad z(0) = z_0 \quad \text{in } \Omega. \tag{13.24}$$

We will study in fact the stabilization of this system around the state $(0, 0)$ for the initial datum $(y_0, z_0)$ lying in a neighborhood of $(0, 0)$. At the end, after proving the stabilization result, by making the backward transformations we obtain the stabilization of the initial system in $(\varphi, \theta)$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

### 13.2.1 Stabilization of the linearized system

We first study the linearized system extracted from (13.22)–(13.24),

$$y_t + \frac{\nu}{\tau^2}(A + A^{-1} - 2)y - \frac{1}{\tau}(A^{-1} - I)(F''(\varphi_\infty)y) \tag{13.25}$$
$$+ \frac{\gamma}{\tau}(A^{-1} - I)z - \frac{\gamma l}{\tau}(A^{-1} - I)y = f_\omega v \quad \text{in } (0,\infty)\times\Omega,$$
$$z_t + \frac{1}{\tau}(A - I)z + \frac{l}{\tau}(I - A)y = f_\omega u \quad \text{in } (0,\infty)\times\Omega, \tag{13.26}$$
$$y(0) = y_0, \quad z(0) = z_0 \quad \text{in } \Omega, \tag{13.27}$$

which can be rewritten in the abstract form

$$\frac{d}{dt}\big(y(t), z(t)\big) + \mathcal{A}\big(y(t), z(t)\big) = f_\omega U(t), \quad \text{a. e. } t \in (0,\infty), \tag{13.28}$$
$$\big(y(0), z(0)\big) = (y_0, z_0), \tag{13.29}$$

where $U(t) = (v(t), u(t))$, with $\mathcal{A}$ is defined on $D(\mathcal{A}) \subset H\times H \to H\times H$ by

$$\mathcal{A} = \begin{bmatrix} \frac{\nu}{\tau^2}(A + A^{-1} - 2) - \frac{\gamma l}{\tau}(A^{-1} - I) - \frac{1}{\tau}(A^{-1} - I)(F''_\infty \cdot) & \frac{\gamma}{\tau}(A^{-1} - I) \\ \frac{l}{\tau}(I - A) & \frac{1}{\tau}(A - I) \end{bmatrix}, \tag{13.30}$$
$$D(\mathcal{A}) = \left\{ w = (y,z) \in L^2(\Omega)\times L^2(\Omega);\ \mathcal{A}w \in \mathcal{H},\ \frac{\partial y}{\partial \nu} = \frac{\partial z}{\partial \nu} = 0 \text{ on } \partial\Omega \right\},$$

where $F''_\infty$ denotes $F''(\varphi_\infty)$.

We set $\mathcal{H} = H\times H$, $\mathcal{V} = D(A^{1/2})\times D(A^{1/2})$, and $\mathcal{V}' = (D(A^{1/2})\times D(A^{1/2}))'$ and note that $\mathcal{V}\subset\mathcal{H}\subset\mathcal{V}'$ algebraically and topologically, with compact injections. The scalar products on $\mathcal{H}$ and $\mathcal{V}$ are defined as

$$\big((y,z),(\psi_1,\psi_2)\big)_{\mathcal{H}} = \int_\Omega \left(\frac{\tau l^2}{\nu} y\psi_1 + z\psi_2\right) dx,$$
$$\big((y,z),(\psi_1,\psi_2)\big)_{\mathcal{V}} = \int_\Omega \left(\frac{\tau l^2}{\nu}(\nabla y\cdot\nabla\psi_1 + y\psi_1) + \nabla z\cdot\nabla\psi_2 + z\psi_2\right) dx.$$

The next result, given without proof because it involves arguments relying on general known theorems, provides the well-posedness of the Cauchy problem (13.28)–(13.29) and some properties of $\mathcal{A}$.

**Proposition 13.2.1.** *The operator $\mathcal{A}$ is quasi m-accretive on $\mathcal{H}$, and its resolvent is compact. Moreover, $-\mathcal{A}$ generates a $C_0$-analytic semigroup. Let $(y_0,z_0)\in\mathcal{H}$ and $(v,u)\in L^2(0,T;\mathcal{H})$. Then problem* (13.28)–(13.29) *has a unique solution $(y,z)\in C([0,T];\mathcal{H})\cap L^2(0,T;\mathcal{V})\cap W^{1,2}(0,T;\mathcal{V}')\cap C((0,T];\mathcal{V})$ for all $T>0$. The norms of the solution in these spaces are bounded by the sum of the norm of the initial data in $\mathcal{H}$ and*

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$\int_0^T \|f_\omega U(s)\|_{\mathcal{H}}^2 ds$, *multiplied by a constant* $C_\infty$ *depending on* $\Omega$, $T$, *the problem parameters, and* $\|F''(\varphi_\infty)\|_\infty$.

In this case the eigenvalues $\lambda_i$ and eigenfunctions $\{(\varphi_i, \psi_i)\}_{i\geq 1}$ of $\mathcal{A}$ are complex. Since the resolvent of $\mathcal{A}$ is compact, there exists a finite number of eigenvalues with real nonpositive parts, $\operatorname{Re}\lambda_i \leq 0$. Each of them may have the order of multiplicity $l_i$, $i = 1, \dots, p$. We write the sequence as $\operatorname{Re}\lambda_1 \leq \operatorname{Re}\lambda_2 \leq \cdots \leq \operatorname{Re}\lambda_N \leq 0$, where each eigenvalue is counted according to its multiplicity, and $N = l_1 + l_2 + \cdots + l_p$. We denote by $\overline{\lambda_i}$ (complex conjugate) and $\{(\varphi_i^*, \psi_i^*)\}_{i\geq 1}$ the eigenvalues and eigenfunctions of the adjoint $\mathcal{A}^*$ of $\mathcal{A}$, that is, $\mathcal{A}^*(\varphi_i^*, \psi_i^*) = \overline{\lambda_i}(\varphi_i^*, \psi_i^*)$, $i \geq 1$, where

$$\mathcal{A}^* = \begin{bmatrix} \frac{\nu}{\tau^2}(A + A^{-1} - 2) - \frac{\gamma l}{\tau}(A^{-1} - I) - \frac{1}{\tau}(A^{-1} - I)(F''(\varphi_\infty)\cdot) & \frac{l}{\tau}(I - A) \\ \frac{\gamma}{\tau}(A^{-1} - I) & \frac{1}{\tau}(A - I) \end{bmatrix}. \tag{13.31}$$

The controller stabilizing the linear system is constructed as a linear combination of the unstable eigenvectors of the adjoint operator $\mathcal{A}^*$ (see, e. g., [6, 21]),

$$f_\omega U(t,x) = \sum_{j=1}^{N} f_\omega \operatorname{Re}\big(\widetilde{w_j}(t)(\varphi_j^*(x), \psi_j^*(x))\big), \quad t \geq 0,\ x \in \Omega, \tag{13.32}$$

where $\widetilde{w_j} \in C([0,\infty); \mathbb{C})$, $j = 1, \dots, N$. This form replaced in (13.28) provides the *open-loop linear system*

$$\begin{aligned} \frac{d}{dt}\big(y(t), z(t)\big) + \mathcal{A}\big(y(t), z(t)\big) &= \sum_{j=1}^{N} f_\omega \operatorname{Re}\big(\widetilde{w_j}(t)(\varphi_j^*(x), \psi_j^*(x))\big), \quad t \in (0,\infty), \\ \big(y(0), z(0)\big) &= \big(y^0, z^0\big), \end{aligned} \tag{13.33}$$

where the initial condition is arbitrary.

**Proposition 13.2.2.** *Let the eigenvalues* $\lambda_i$ *be semisimple and assume that* $\varphi_\infty$ *is an analytic function in* $\Omega$. *Then there exist* $w_j \in L^2(\mathbb{R}^+)$, $j = 1, \dots, 2N$, *such that controller* (13.32) *stabilizes exponentially system* (13.33), *that is, its solution* $(y, z)$ *satisfies*

$$\|y(t)\|_H + \|z(t)\|_H \leq C_\infty e^{-k_\infty t}\big(\|y^0\|_H + \|z^0\|_H\big) \quad \textit{for all } t \geq 0. \tag{13.34}$$

*Moreover, we have*

$$\left(\sum_{j=1}^{2N} \int_0^\infty |w_j(t)|^2 dt\right)^{1/2} \leq C\big(\|y^0\|_H + \|z^0\|_H\big), \tag{13.35}$$

*where* $C_\infty$, $C$, $k_\infty$ *depend on the problem parameters* $\nu$, $\gamma$, $l$ *and on* $\Omega$ *and* $\|F''(\varphi_\infty)\|_\infty$.

*Proof.* (Sketch) Since the eigenfunctions are complex, we must work in the compexified space $\widetilde{\mathcal{H}} = \mathcal{H} + i\mathcal{H}$, $i = \sqrt{-1}$, to stabilize the following system for the complex

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

functions $(\widetilde{y},\widetilde{z}) = (y,z) + i(Y,Z)$:

$$\frac{d}{dt}(\widetilde{y}(t),\widetilde{z}(t)) + \mathcal{A}(\widetilde{y}(t),\widetilde{z}(t)) = \sum_{j=1}^{N} f_\omega(\widetilde{w_j}(t)(\varphi_j^*(x),\psi_j^*(x)), \quad \text{a. e. } t \in (0,\infty), \tag{13.36}$$

$$(\widetilde{y}(0),\widetilde{z}(0)) = (y^0,z^0). \tag{13.37}$$

Finally, we take $(y,z) = (\operatorname{Re}\widetilde{y},\operatorname{Re}\widetilde{z})$, $(v,u) = (\operatorname{Re}\widetilde{v},\operatorname{Re}\widetilde{u})$, and the pair $(y,z)$ constructed in this way turns out to be a solution to the open-loop system (13.33) corresponding to controller (13.32). We consider the linear space generated by the eigenfunctions $\{(\varphi_i,\psi_i)\}_{i=1,\dots,N}$ and denote it by $\widetilde{\mathcal{H}_N} := \operatorname{lin\,span}\{(\varphi_1,\psi_1),\dots,(\varphi_N,\psi_N)\}$. Also, $\widetilde{\mathcal{H}_S} := \operatorname{lin\,span}\{(\varphi_{N+1},\psi_{N+1}),\dots\}$. We have the algebraic decomposition $\widetilde{\mathcal{H}} = \widetilde{\mathcal{H}_N} \oplus \widetilde{\mathcal{H}_S}$ unique but not orthogonal. Moreover, $\mathcal{A}_N = \mathcal{A}|_{\widetilde{\mathcal{H}_N}}$ has the eigenfunctions $\{\varphi_i,\psi_i\}_{i=1,\dots,N}$, and $\mathcal{A}_S = \mathcal{A}|_{\widetilde{\mathcal{H}_S}} = (I - P_N)\mathcal{A}$ has the eigenfunctions $\{\varphi_i,\psi_i\}_{i\geq N+1}$, where $P_N : \mathcal{H} \to \mathcal{H}_N$ is the algebraic projector. On the invariant subspace $\widetilde{\mathcal{H}_S}$ the operator $-\mathcal{A}_S$ generates a $C_0$-analytic semigroup, that is, $\|e^{-\mathcal{A}_S t}\|_{\mathcal{L}(\widetilde{\mathcal{H}_S}\times\widetilde{\mathcal{H}_S})} \leq Ce^{-\widetilde{k}t}$, $\widetilde{k} = \operatorname{Re}(\lambda_{N+1} - \lambda_N)$. Now, system (13.36)–(13.37) is split into

$$\frac{d}{dt}(y_N(t),z_N(t)) + \mathcal{A}_N(y_N(t),z_N(t)) = P_N\left(\sum_{j=1}^{N} f_\omega \widetilde{w_j}(t)(\varphi_j^*,\psi_j^*)\right),$$

$$(y_N(0),z_N(0)) = P_N(y^0,z^0), \tag{13.38}$$

and

$$\frac{d}{dt}(y_S(t),z_S(t)) + \mathcal{A}_S(y_S(t),z_S(t)) = (I - P_N)\sum_{j=1}^{N} f_\omega \widetilde{w_j}(t)(\varphi_j^*,\psi_j^*),$$

$$(y_S(0),z_S(0)) = (I - P_N)(y^0,z^0). \tag{13.39}$$

Let $T_0 > 0$ be arbitrary, fixed. We will prove that system (13.38) is null controllable in $T_0$ and the solution to (13.39) decreases exponentially to 0 as $t \to \infty$. The solution to (13.36)–(13.37) is represented as

$$(\widetilde{y}(t,x),\widetilde{z}(t,x)) = \sum_{j=1}^{\infty} \xi_j(t)(\varphi_j(x),\psi_j(x)), \quad (t,x) \in (0,\infty)\times\Omega, \tag{13.40}$$

with $\xi_j \in C([0,\infty);\mathbb{C})$, and since $\lambda_i$ are semisimple, the system $\{(\varphi_i,\psi_i)\}_i$, $\{(\varphi_i^*,\psi_i^*)\}_i$ is biorthogonal, that is, $((\varphi_i,\psi_i),(\overline{\varphi_j^*},\overline{\psi_j^*}))_{H\times H} = \delta_{ij}$. Replacing (13.40) in the system and multiplying the equation scalarly by $(\overline{\varphi_i^*},\overline{\psi_i^*})$, we get

$$\xi_i' + \lambda_i\xi_i = \sum_{j=1}^{N} \widetilde{w_j} d_{ij}, \quad \xi_i(0) = \xi_{i0} \quad \text{for } i \geq 1, \tag{13.41}$$

$$\xi_i(0) = \xi_{i0},$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

where

$$d_{ij} = \int_\Omega f_\omega(\varphi_i^* \overline{\varphi_j^*} + \psi_i^* \overline{\psi_j^*})dx, \quad j = 1, \dots, N,\ i \geq 1. \tag{13.42}$$

The values $|d_{ij}| \leq C_\infty$, where $C_\infty$ depends on problem data and $\|F''(\varphi_\infty)\|_\infty$ and reduces to $\|\varphi_\infty\|_\infty$ in the case with polynomial potential.

We can be prove by using the Kalman lemma and the fact that the system $\{\sqrt{f_\omega}\varphi_j^*, \sqrt{f_\omega}\psi_j^*\}_{j=1}^N$ is linearly independent on $\omega$ (based on the analyticity of $\varphi_\infty$ in $\Omega$) that (13.41) for $i = 1, \dots, N$ is null controllable in $T_0 > 0$, that is,

$$\xi_i(T_0) = 0 \quad \text{and} \quad \xi_i(t) = 0 \quad \text{for } t > T_0,\ i = 1, \dots, N.$$

We now treat system (13.41) for $i > N + 1$. By using the variation-of-constants formula and appropriate estimates for the solution provided by it we get that system (13.41) is exponentially stabilized in the origin. Hence (13.38) is null controllable in $T_0$.

For system (13.39), we show the exponential stabilization in the origin relying on the properties of the $C_0$-analytic semigroup $-\mathcal{A}_S$.

These findings imply that

$$\|(\widetilde{y}, \widetilde{z})(t)\|_{\widetilde{\mathcal{H}}} \leq Ce^{-kt}\|(y^0, z^0)\|_{\widetilde{\mathcal{H}}} \quad \text{for } t > 0 \tag{13.43}$$

and $\|(\widetilde{y}, \widetilde{z})(t)\|_{\widetilde{\mathcal{H}}} \to 0$ as $t \to \infty$.

Finally, we find that the controller $(v, u)$ involves a sequence of $2N$ terms obtained by setting

$$w_j := \operatorname{Re} \widetilde{w_j} \quad \text{for } j = 1, \dots, N, \quad w_{j+N} := \operatorname{Im} \widetilde{w_j} \quad \text{for } j = 1, \dots, N \tag{13.44}$$

and can be expressed as

$$\begin{aligned} v(t,x) &= \sum_{j=1}^N w_j(t) \operatorname{Re} \varphi_j^*(x) - \sum_{j=1}^N w_{j+N}(t) \operatorname{Im}\ \varphi_j^*(x), \\ u(t,x) &= \sum_{j=1}^N w_j(t) \operatorname{Re} \psi_j^*(x) - \sum_{j=1}^N w_{j+N}(t) \operatorname{Im}\ \psi_j^*(x). \end{aligned} \tag{13.45}$$

Then we take $(y, z) = (\operatorname{Re} \widetilde{y}, \operatorname{Re} \widetilde{z})$, and we get by (13.43) the stabilization inequality (13.34), as claimed. □

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

### 13.2.2 Construction of the feedback controller

We will find the feedback controller (depending on the solution $(y,z)$), which exponentially stabilizes the solution to (13.33) by solving the minimization problem

$$\Phi(y^0,z^0)=\min_{W\in L^2(0,\infty;\mathbb{R}^{2N})}\left\{\frac{1}{2}\int_0^\infty(\|Ay(t)\|_H^2+\|Az(t)\|_H^2+\|W(t)\|_{\mathbb{R}^{2N}}^2)dt\right\} \tag{13.46}$$

subject to (13.33). Here $W=(w_1,\dots,w_N,w_{N+1},\dots,w_{2N})\in L^2(0,\infty;\mathbb{R}^{2N})$ is defined in (13.44). We note that $D(\Phi)=\{(y^0,z^0)\in H\times H;\Phi(y^0,z^0)<\infty\}$. Let $\mathbb{R}^+=(0,+\infty)$.

**Proposition 13.2.3.** *For each pair $(y^0,z^0)\in D(A^{1/2})\times D(A^{1/2})$, problem* (13.46) *has a unique optimal solution*

$$(\{w_j^*\}_{j=1}^{2N},y^*,z^*)\in L^2(\mathbb{R}^+;\mathbb{R}^{2N})\times L^2(\mathbb{R}^+;D(A^{1/2}))\times L^2(\mathbb{R}^+;D(A^{1/2})), \tag{13.47}$$

*which satisfies*

$$\begin{aligned}c_1(\|A^{1/2}y^0\|_H^2+\|A^{1/2}z^0\|_H^2)&\le\Phi(y^0,z^0)\\&\le c_2(\|A^{1/2}y^0\|_H^2+\|A^{1/2}z^0\|_H^2).\end{aligned} \tag{13.48}$$

*If $(y^0,z^0)\in D(A)\times D(A)$, then*

$$\begin{aligned}&(\|Ay^*(t)\|_H^2+\|Az^*(t)\|_H^2)+\int_0^t(\|A^{3/2}y^*(s)\|_H^2+\|A^{3/2}z^*(s)\|_H^2)ds\\&\le c_3(\|Ay^0\|_H^2+\|A^{1/2}z^0\|_H^2)\quad\text{for all } t\ge 0,\end{aligned} \tag{13.49}$$

*where $c_1$, $c_2$, $c_3$ are positive constants (depending on $\Omega$, the problem parameters, and $\|F''(\varphi_\infty)\|_\infty$.)*

*Proof.* (Idea) The proof is led in two steps. First, we prove that (13.46) has a solution by using the standard technique involving a minimization sequence. The solution is unique because the functional is strictly convex and the state system is linear. Then appropriate technical estimates of the norms of the solution to (13.33) lead to properties (13.48)–(13.49). Here we will not present them, because they are very long and are provided in detail in [18]. □

An important consequence of this result is that there exists a functional $R:\mathcal{V}\to\mathcal{V}'$ such that

$$\begin{aligned}\Phi(y^0,z^0)&=\frac{1}{2}\langle R(y^0,z^0),(y^0,z^0)\rangle_{\mathcal{V}',\mathcal{V}}\\&\text{for all }(y^0,z^0)\in\mathcal{V}=D(A^{1/2})\times D(A^{1/2}).\end{aligned} \tag{13.50}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

In particular, $R(y^0, z^0)$ is exactly the Gâteaux derivative of the function $\Phi$ at $(y^0, z^0)$,

$$\Phi'(y^0, z^0) = R(y^0, z^0) \quad \text{for all } (y^0, z^0) \in \mathcal{V}. \tag{13.51}$$

Next, we restrict the domain of $R$ to $D(R) := \{(y^0, z^0) \in \mathcal{V}; R(y^0, z^0) \in H \times H\}$ and still denote by $R$ the restriction $R : D(R) \subset \mathcal{V} \subset H \times H \to H \times H$. This is viewed now as on operator on $H \times H$, and so we call it the restriction on $H \times H$. Since $\Phi$ is coercive by (13.48), it follows that $R$ is self-adjoint.

Let us introduce now the operators $B : \mathbb{R}^{2N} \to H \times H$ and $B^* : H \times H \to \mathbb{R}^{2N}$ by

$$\begin{aligned} Bp &= \begin{bmatrix} f_\omega(\sum_{j=1}^N p_j \operatorname{Re} \varphi_j^* - \sum_{j=N+1}^{2N} p_j \operatorname{Im} \varphi_j^*) \\ f_\omega(\sum_{j=1}^N p_j \operatorname{Re} \psi_j^* - \sum_{j=N+1}^{2N} p_j \operatorname{Im} \psi_j^*) \end{bmatrix} \\ &\text{for } p = \begin{bmatrix} p_1 \\ \cdots \\ p_{2N} \end{bmatrix} \in \mathbb{R}^{2N} \end{aligned} \tag{13.52}$$

and

$$\begin{aligned} B^* q &= \begin{bmatrix} \int_\Omega f_\omega(q_1 \operatorname{Re} \varphi_1^* + q_2 \operatorname{Re} \psi_1^*) dx \\ \cdots \\ \int_\Omega f_\omega(q_1 \operatorname{Re} \varphi_N^* + q_2 \operatorname{Re} \psi_N^*) dx - \int_\Omega f_\omega(q_1 \operatorname{Im} \varphi_1^* + q_2 \operatorname{Im} \psi_1^*) dx \\ \cdots \\ - \int_\Omega f_\omega(q_1 \operatorname{Im} \varphi_N^* + q_2 \operatorname{Im} \psi_N^*) dx \end{bmatrix} \\ &\text{for all } q = \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} \in H \times H. \end{aligned} \tag{13.53}$$

Then (13.33) can be rewritten as

$$\begin{aligned} \frac{d}{dt}(y(t), z(t)) + \mathcal{A}(y(t), z(t)) &= BW(t), \quad \text{a. e. } t > 0, \\ (y(0), z(0)) &= (y^0, z^0). \end{aligned} \tag{13.54}$$

In the next proposition, we provide the feedback representation for the optimal solution to (13.46).

**Proposition 13.2.4.** *Let $W^* = \{w_i^*\}_{i=1}^{2N}$ and $(y^*, z^*)$ be optimal for problem* (13.46)*, corresponding to $(y^0, z^0) \in D(A^{1/2}) \times D(A^{1/2})$. Then $W^*$ is expressed as*

$$W^*(t) = -B^* R(y^*(t), z^*(t)) \quad \textit{for all } t > 0. \tag{13.55}$$

*Moreover, $R$ has the following properties:*

$$\begin{aligned} &2c_1 \|(y^0, z^0)\|_{\mathcal{V}}^2 \le \langle R(y^0, z^0), (y^0, z^0) \rangle_{\mathcal{V}' \times \mathcal{V}} \le 2c_2 \|(y^0, z^0)\|_{\mathcal{V}}^2 \\ &\quad \textit{for all } (y^0, z^0) \in \mathcal{V} = D(A^{1/2}) \times D(A^{1/2}), \end{aligned} \tag{13.56}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$$\|R(y^0,z^0)\|_{H\times H} \le C_R\|(y^0,z^0)\|_{D(A)\times D(A)} \quad \textit{for all } (y^0,z^0)\in D(A)\times D(A), \tag{13.57}$$

*and $R$ satisfies the Riccati algebraic equation*

$$\begin{aligned} &2\big(R(\overline{y},\overline{z}),\mathcal{A}(\overline{y},\overline{z})\big)_{H\times H} + \|B^*R(\overline{y},\overline{z})\|^2_{\mathbb{R}^{2N}} \\ &\quad = \|A\overline{y}\|^2_H + \|Az\|^2_H \quad \textit{for all } (\overline{y},\overline{z})\in D(A)\times D(A). \end{aligned} \tag{13.58}$$

*Here $c_1$, $c_2$, $C_R$ are constants depending on the problem parameters, $\Omega$, and $\|F''(\varphi_\infty)\|_\infty$.*

*Proof.* For the reader's convenience, we indicate the steps in this proof. Let $T$ be positive and arbitrary. We introduce the minimization problem

$$\mathrm{Min}\left\{\frac{1}{2}\int_0^T\big(\|Ay(t)\|^2_H + \|Az(t)\|^2_H + \|W(t)\|^2_{\mathbb{R}^{2N}}\big)dt + \Phi(y(T),z(T))\right\} \tag{13.59}$$

for $W\in L^2(0,T;\mathbb{R}^{2N})$ subject to (13.33). This is equivalent to (13.46) by the dynamic programming principle (see, e. g., [1], p. 104). We introduce the adjoint system

$$\begin{aligned} &\frac{d}{dt}(p^T,q^T)(t) - \mathcal{A}^*(p^T(t),q^T(t)) = (A^2y^*(t),A^2z^*(t)) \quad \text{in } (0,T)\times\Omega, \\ &\qquad (p^T(T),q^T(T)) = -R(y^*(T),z^*(T)) \quad \text{in } \Omega, \end{aligned} \tag{13.60}$$

where the final condition at $t=T$ follows by (13.51). As we shall see, the solution to (13.60) is independent of $T$. By the maximum principle in (13.59), we have

$$W^*(t) = B^*\big(p^T(t),q^T(t)\big), \quad \text{a. e. } t\in(0,T) \tag{13.61}$$

(see [16], p. 114; see also [1], p. 190).

For proving (13.57), let $(y^0,z^0)\in D(A)\times D(A)$. First, we show that $(p^T,q^T)$ is in $C([0,T);H\times H)$ by an adapted argument following the idea from [7] and [4]. We define $(\widetilde{p},\widetilde{q}) = \widetilde{A}(p^T,q^T)$, where $\widetilde{A}$ is the operator

$$\widetilde{A} = \begin{bmatrix} A^{-1/2} & 0 \\ 0 & A^{-1/2} \end{bmatrix}.$$

By recalling (13.31) we see that $\mathcal{A}^*$ and $\widetilde{A}$ commute, and so we obtain the system

$$\begin{aligned} &\frac{d}{dt}(\widetilde{p},\widetilde{q})(t) - \mathcal{A}^*(\widetilde{p}(t),\widetilde{q}(t)) = (A^{3/2}y^*(t),A^{3/2}z^*(t)) \quad \text{in } (0,T)\times\Omega, \\ &\qquad (\widetilde{p}(T),\widetilde{q}(T)) = -\widetilde{A}R(y^*(T),z^*(T)) \quad \text{in } \Omega. \end{aligned} \tag{13.62}$$

By (13.49) we have $(A^{3/2}y^*,A^{3/2}z^*)\in L^2(0,T;H\times H)$. Since $R(y^*(T),z^*(T))\in V'\times V'$, we get $\widetilde{A}R(y^*(T),z^*(T))\in H\times H$. By applying a backward version of Proposition 13.2.1

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

we see that system (13.62) has a unique solution $(\widetilde{p},\widetilde{q}) \in C([0,T);\mathcal{V})$, and so $(p^T,q^T) \in C([0,T);H\times H)$. Next, we prove the relation

$$R(y^0,z^0) = -(p^T(0),q^T(0)). \tag{13.63}$$

To this end, let us consider two solutions to (13.59), $(W^*,y^*,z^*)$ and $(W_1^*,y_1^*,z_1^*)$, corresponding to $(y^0,z^0)$ and $(y^1,z^1)$, respectively, both belonging to $D(A)\times D(A)$, and show the relation

$$\Phi(y^0,z^0) - \Phi(y^1,z^1) \le -((p^T(0),q^T(0)),(y^0-y^1,z^0-z^1))_{H\times H}. \tag{13.64}$$

This implies that $-(p^T(0),q^T(0)) \in \partial\Phi(y^0,z^0)$. Since $\Phi$ is differentiable on $D(A^{1/2})\times D(A^{1/2})$, it follows that $-(p^T(0),q^T(0)) = \Phi'(y^0,z^0) = R(y^0,z^0)$, as claimed in (13.63). Since $(p^T,q^T) \in C([0,\infty);H\times H)$, we get $(p^T(0),q^T(0)) \in H\times H$, and so $R(y^0,z^0) \in H\times H$ for all $(y^0,z^0) \in D(A)\times D(A)$. Since $R$ is a linear closed operator from $D(A)\times D(A)$ to $H\times H$, by the closed graph theorem we conclude that it is continuous (see, e. g., [9], Thm. 2.9, p. 37), that is, $R \in \mathcal{L}(D(A)\times D(A);H\times H)$, as claimed by (13.57).

We define the restriction of $R$ to $H\times H$, still denoted by $R$, in the sense specified in Proposition 13.2.3. Thus its domain contains $D(A)\times D(A)$. Next, resuming (13.61), which extends by continuity at $t=T$, in $V'$, we get

$$W^*(T) = B^*(p^T(T),q^T(T)). \tag{13.65}$$

Moreover, since $(y^*(t),z^*(t)) \in D(A)\times D(A)$ for all $t \ge 0$, by (13.49) we have that $R(y^*(t),z^*(t)) \in H\times H$ for all $t \ge 0$. This is true also for $t=T$, and so using the final condition in (13.60), we get

$$(p^T(T),q^T(T)) = -R(y^*(T),z^*(T)) \in H\times H. \tag{13.66}$$

This relation, combined with (13.65), implies $W^*(T) = -B^*R(y^*(T),z^*(T))$, where $T$ is arbitrary. Therefore (13.55) follows.

Inequalities (13.56) follow immediately by (13.50) and (13.48). By (13.55) we also remark that

$$f_\omega U(t) = f_\omega(v^*(t),u^*(t)) = -BB^*R(y^*(t),z^*(t)), \tag{13.67}$$

which can be used to give expressions of $u^*$ and $v^*$.

To prove (13.58), we consider $(y^0,z^0) \in D(A)\times D(A)$. By (13.46) and (13.59) written with $T=t$ we get

$$\Phi(y^*(t),z^*(t)) = \frac{1}{2}\int_t^\infty (\|Ay^*(s)\|_H^2 + \|Az^*(s)\|_H^2 + \|W^*(s)\|_{\mathbb{R}^{2N}}^2)ds \tag{13.68}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

for any $t \geq 0$. We note that

$$\begin{aligned}\left\|BB^*R(y^*(t),z^*(t))\right\|_{H\times H} &\leq C_1\left\|R(y^*(t),z^*(t))\right\|_{H\times H}\\ &\leq C_2\left\|(y^*(t),z^*(t))\right\|_{D(A)\times D(A)},\end{aligned}$$

since $(Ay^*(t),Az^*(t)) \in H\times H$ and $R(y^*(t),z^*(t)) \in \mathcal{H}$ for a. e. $t>0$.

System (13.54) in which the right-hand side is replaced by (13.67) becomes a *closed-loop* system with the right-hand side $-BB^*R(y^*(t),z^*(t))$. It can be shown that $-(\mathcal{A}+BB^*R)$ generates a $C_0$-semigroup on $H\times H$ (see Lemma A3 in [4], Appendix). Hence the closed-loop system (13.54) has, for $(y^0,z^0)\in D(A)\times D(A)=D(\mathcal{A})$, a unique weak solution $(y^*(t),z^*(t))\in C([0,\infty);H\times H)$ (see [2], p. 141) such that

$$\mathcal{A}(y^*(t),z^*(t))+BB^*R(y^*(t),z^*(t))+\frac{d}{dt}(y^*(t),z^*(t))\in L^\infty(0,\infty;\mathcal{H}).$$

Since $BB^*R(y^*(t),z^*(t))\in L^2(0,\infty;\mathcal{H})$, we have $\mathcal{A}(y^*(t),z^*(t))\in L^2(0,\infty;\mathcal{H})$.

Now we differentiate (13.68) with respect to $t$, recalling (13.50) and that $R$ is symmetric. We get

$$\begin{aligned}&\left(R(y^*(t),z^*(t)),\frac{d}{dt}(y^*(t),z^*(t))\right)_{H\times H}+\frac{1}{2}\left(\left\|Ay^*(t)\right\|_H^2+\left\|Az^*(t)\right\|_H^2\right)\\ &\quad+\frac{1}{2}\left\|B^*R(y^*(t),z^*(t))\right\|_{\mathbb{R}^{2N}}^2=0,\quad \text{a. e. } t>0.\end{aligned}\tag{13.69}$$

Then replacing $\frac{d}{dt}(y^*(t),z^*(t))$ from (13.54) in (13.69) and taking into account (13.55), we get

$$\begin{aligned}&\left(R(y^*(t),z^*(t)),-\mathcal{A}(y^*(t),z^*(t))\right)_{H\times H}+\frac{1}{2}\left(\left\|Ay^*(t)\right\|_H^2+\left\|Az^*(t)\right\|_H^2\right)\\ &\quad+\frac{1}{2}\left\|B^*R(y^*(t),z^*(t))\right\|_{\mathbb{R}^{2N}}^2=\left(R(y^*(t),z^*(t)),BB^*R(y^*(t),z^*(t))\right)_{H\times H}\end{aligned}$$

for $t\geq 0$, which implies (13.58). □

## 13.2.3 Feedback stabilization of the nonlinear system

The last part of the proof concerns the nonlinear system (13.22)–(13.24) in which the right-hand side $(f_\omega v, f_\omega u)$ is replaced by the feedback controller determined in the previous section, that is, by $f_\omega U(t)=-BB^*R(y(t),z(t))$. We consider the closed-loop system in the abstract form

$$\begin{aligned}&\frac{d}{dt}(y(t),z(t))+\mathcal{A}(y(t),z(t))=\mathcal{G}(y(t))-BB^*R(y(t),z(t)),\quad \text{a. e. } t>0,\\ &\qquad(y(0),z(0))=(y_0,z_0),\end{aligned}\tag{13.70}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

where $\mathcal{G}(y(t)) = (G(y(t)), 0)$ and $G(y) = \frac{1}{\tau}(A^{-1} - I)F_r(y)$. We recall that $F_r$ is the second-order remainder of the Taylor expansion of $F'(y+\varphi_\infty)$, which we express for simplicity in the integral form

$$F_r(y) = y^2 \int_0^1 (1-s)F'''(\varphi_\infty + sy)dy = y^3 + 3\varphi_\infty y^2. \tag{13.71}$$

Next, we prove a stabilization result.

**Theorem 13.2.5.** *Let $(y_0, z_0) \in D(A^{1/2}) \times D(A^{1/2})$. There exists $\rho$ such that if*

$$\|y_0\|_{D(A^{1/2})} + \|z_0\|_{D(A^{1/2})} \le \rho, \tag{13.72}$$

*then the closed-loop system* (13.70) *has a unique solution*

$$\begin{aligned}(y,z) \in C\big([0,\infty); H\times H\big) \cap L^2\big(0,\infty; D(A)\times D(A)\big)\\ \cap W^{1,2}\big(0,\infty; \big(D(A^{1/2})\times D(A^{1/2})\big)'\big),\end{aligned} \tag{13.73}$$

*which is exponentially stable:*

$$\big\|y(t)\big\|_{D(A^{1/2})} + \big\|z(t)\big\|_{D(A^{1/2})} \le C_\infty e^{-k_\infty t}\big(\|y_0\|_{D(A^{1/2})} + \|z_0\|_{D(A^{1/2})}\big) \tag{13.74}$$

*for some positive constants $k_\infty$ and $C_\infty$, which depend on $\Omega$, the problem parameters, and $\|\varphi_\infty\|_\infty$.*

*Proof.* In the first step, we prove existence and uniqueness for (13.70). These will be shown first on every interval $[0,T]$ by the Schauder fixed point theorem, and after that these extend to $[0,\infty)$ by the unique solution continuity. Let $r$ be a positive constant, which will be specified later. For arbitrary fixed $T > 0$, we introduce the set

$$\begin{aligned}S_T = \Big\{ &(y,z) \in L^2(0,T; H\times H);\ \sup_{t\in(0,T)} \big(\big\|y(t)\big\|^2_{D(A^{1/2})} + \big\|z(t)\big\|^2_{D(A^{1/2})}\big)\\ &+ \int_0^T \big(\big\|Ay(t)\big\|_H^2 + \big\|Az(t)\big\|_H^2\big)dt \le r^2\Big\},\end{aligned} \tag{13.75}$$

which is a convex closed subset of $L^2(0,T; D(A^{1/2})\times D(A^{1/2}))$.

We fix $(\overline{y}, \overline{z}) \in S_T$, consider the Cauchy problem

$$\begin{aligned}\frac{d}{dt}\big(y(t), z(t)\big) + \mathcal{A}\big(y(t), z(t)\big) + BB^*R\big(y(t), z(t)\big) = \mathcal{G}\big(\overline{y}(t)\big), \quad \text{a. e. } t > 0,\\ \big(y(0), z(0)\big) = (y_0, z_0),\end{aligned} \tag{13.76}$$

and prove that the solution to this problem exists and is unique.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Then we define $\Psi_T : S_T \to L^2(0,T;\mathcal{V})$ by $\Psi_T(\overline{y},\overline{z}) = (y,z)$ the solution to (13.76) and show that this mapping has the following properties: (a) $\Psi_T(S_T) \subset S_T$ for a suitable $r$; (b) $\Psi_T(S_T)$ is relatively compact in $L^2(0,T;D(A^{1/2}) \times D(A^{1/2}))$; and (c) $\Psi_T$ is continuous in the $L^2(0,T;D(A^{1/2}) \times D(A^{1/2}))$ norm.

A detailed proof that ensures that problem (13.76) has a unique solution

$$(y,z) \in C([0,T];\mathcal{V}) \cap W^{1,2}(0,T;\mathcal{H}) \cap L^2(0,T;D(\mathcal{A})), \tag{13.77}$$

provided that $\mathcal{G}(\overline{y}) \in L^2(0,T;H \times H)$, can be done by following [4], Theorem 3.1. Moreover, relation (13.77) implies that $(y(t),z(t)) \in D(A) \times D(A)$ for a. e. $t \in (0,T)$, and so $R(y(t),z(t)) \in H \times H$ for a. e. $t \in (0,T)$.

It remains to show that $\mathcal{G}(\overline{y}) \in L^2(0,T;H \times H)$. To this end, we calculate

$$\begin{aligned} \|G(\overline{y}(t))\|_H^2 &= \left\| \frac{1}{\tau}(A^{-1} - I)F_r(\overline{y}(t)) \right\|_H^2 \le C_G \|F_r(\overline{y}(t))\|_H^2 \\ &\le C_G \|\overline{y}^3(s) + 3\varphi_\infty \overline{y}^2(s)\|_H^2 \le C_G(\|\overline{y}(t)\|_{D(A^{1/2})}^6 + \|\varphi_\infty\|_\infty^2 \|\overline{y}(t)\|_{D(A^{1/2})}^4) \end{aligned} \tag{13.78}$$

with $C_G$ depending on $\Omega$ and the problem parameters. Since $(\overline{y},\overline{z}) \in S_T$, it follows

$$\begin{aligned} \int_0^t \|G(\overline{y}(s))\|_H^2 ds &\le C_G(r^4 + \|\varphi_\infty\|_\infty^2 r^2) \int_0^t \|A\overline{y}(s)\|_H^2 ds \\ &\le C_G(r^6 + \|\varphi_\infty\|_\infty^2 r^4). \end{aligned} \tag{13.79}$$

To prove that $(y,z) \in S_T$, we multiply (13.76) by $R(y(t),z(t)) \in \mathcal{H}$ scalarly in $\mathcal{H}$,

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}(R(y(t),z(t)),(y(t),z(t)))_{H\times H} + (\mathcal{A}(y(t),z(t)),R(y(t),z(t)))_{H\times H} \\ &\quad = -\|B^*R(y(t),z(t))\|_{\mathbb{R}^{2N}}^2 + (\mathcal{G}(\overline{y}(t)),R(y(t),z(t)))_{H\times H}, \quad \text{a. e. } t > 0, \end{aligned}$$

and then use the Riccati equation (13.58). Recalling (13.57), we obtain

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}(R(y(t),z(t)),(y(t),z(t)))_{H\times H} \\ &\qquad + \frac{1}{2}(\|Ay(t)\|_H^2 + \|Az(t)\|_H^2 + \|B^*R(y(t),z(t))\|_{\mathbb{R}^{2N}}^2) \\ &\quad \le \frac{1}{4}(\|Ay(t)\|_H^2 + \|Az(t)\|_H^2) + 4C_R^2\|G(\overline{y}(t))\|_H^2, \quad \text{a. e. } t \in (0,T). \end{aligned}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Integrating over $(0,t)$ and using (13.56), we finally get

$$\|y(t)\|^2_{D(A^{1/2})} + \|z(t)\|^2_{D(A^{1/2})} + \frac{1}{4c_1}\int_0^t (\|Ay(s)\|^2_H + \|Az(s)\|^2_H)ds$$

$$\le \frac{c_2}{c_1}(\|y_0\|^2_{D(A^{1/2})} + \|z_0\|^2_{D(A^{1/2})}) + \frac{4C_R^2}{c_1}\int_0^t \|G(\overline{y}(s))\|^2_H ds. \tag{13.80}$$

It remains to impose that the right-hand side is less than $r^2$. Using (13.79), we obtain

$$\frac{c_2}{c_1}\rho^2 + \overline{C_1}(r^6 + \|\varphi_\infty\|^2_\infty r^4) \le r^2, \tag{13.81}$$

where $\overline{C_1} = \frac{4C_R^2 C_G}{c_1}$ depends on the problem parameters, $\Omega$, and $\|F(\varphi_\infty)\|_\infty$ (which, in this case, is proportional to $\|\varphi_\infty\|_\infty$). Relation (13.81) is satisfied, e. g., by the choice $\frac{c_2}{c_1}\rho^2 = \frac{r^2}{2}$, that is, by taking

$$\rho := r\sqrt{\frac{c_1}{2c_2}} \tag{13.82}$$

and by fixing $r$ from the inequality $2\overline{C_1}r^4 + 2\overline{C_1}\|\varphi_\infty\|^2_\infty r^2 - 1 \le 0$. This yields

$$0 < r \le r_1 := \sqrt{\frac{-\overline{C_1}\|\varphi_\infty\|^2_\infty + \sqrt{\overline{C_1}^2\|\varphi_\infty\|^4_\infty + 2\overline{C_1}}}{2\overline{C_1}}}. \tag{13.83}$$

(b) The fact that $\Psi_T(S_T)$ is relatively compact in $L^2(0,T;D(A^{1/2})\times D(A^{1/2}))$ follows by (13.77), because $\frac{d}{dt}(y,z) \in L^2(0,T;H\times H)$, $(y,z) \in L^2(0,T;D(A)\times D(A))$ and $D(A)\times D(A)$ is compactly embedded in $D(A^{1/2})\times D(A^{1/2})$.

(c) The last point to be checked is the continuity of $\Psi_T$. Let $(\overline{y_n},\overline{z_n}) \in S_T$ ve such that $(\overline{y_n},\overline{z_n}) \to (\overline{y},\overline{z})$ strongly in $L^2(0,T;D(A^{1/2})\times D(A^{1/2}))$ as $n\to\infty$. We have to prove that the corresponding solution $(y_n,z_n) = \Psi_T(\overline{y_n},\overline{z_n})$ to (13.76) converges strongly to $(y,z) = \Psi_T(\overline{y},\overline{z})$ in $L^2(0,T;D(A^{1/2})\times D(A^{1/2}))$. The solution $(y_n,z_n)$ to (13.76) corresponding to $(\overline{y_n},\overline{z_n})$ is bounded in the spaces (13.77) due to estimate (13.80). Hence, on a subsequence $\{n\to\infty\}$, it follows that

$$(y_n,z_n) \to (y,z) \quad \text{weakly in } L^2(0,T;D(A)\times D(A)),$$
$$\left(\frac{dy_n}{dt},\frac{dz_n}{dt}\right) \to \left(\frac{dy}{dt},\frac{dz}{dt}\right) \quad \text{weakly in } L^2(0,T;H\times H),$$

and by the Aubin–Lions lemma

$$(y_n,z_n) \to (y,z) \quad \text{strongly in } L^2(0,T;D(A^{1/2})\times D(A^{1/2})).$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Because $\{\overline{y_n}(t)\}_n$ is bounded in $V$ for all $t \in [0,T]$, by the Sobolev embedding $V \subset L^6(\Omega) \subset L^4(\Omega)$, which takes place if if $\Omega \subset \mathbb{R}^d$ with $d \in \{1,2,3\}$, it follows that $\{\overline{y_n}^3(t)\}_n$ and $\{\overline{y_n}^2(t)\}_n$ are bounded in $L^2(\Omega)$. Moreover, since $\overline{y_n} \to \overline{y}$ strongly in $L^2(0,T;V)$ and, consequently, a. e. on $(0,T)\times\Omega$, we deduce that $\overline{y_n}^3 \to \overline{y}^3$ and $\overline{y_n}^2 \to \overline{y}^2$ weakly in $L^2(0,T;H)$. These imply that $F_r(\overline{y_n}) \to F_r(\overline{y})$ weakly in $L^2(0,T;H)$ and also $G(\overline{y_n}) \to G(\overline{y})$ weakly in $L^2(0,T;H)$.

Now writing the weak form of (13.76) corresponding to $(\overline{y_n},\overline{z_n})$ and passing to the limit, we get that $(y,z) = \Psi_T(\overline{y},\overline{z})$. As the same holds for any subsequence, this ends the proof of the continuity of $\Psi_T$.

To conclude, $\Psi$ satisfies the conditions of Schauder's theorem and has a fixed point $y$, $\Psi(y) = y$.

The uniqueness is proved using system (13.22)–(13.24). To this end, we apply a standard technique by estimating the difference of two solutions by using very fine estimates. The whole proof can be found in [4] and [18].

The second step is to prove the stabilization result. To this end, we multiply equation (13.70) by $R(y(t),z(t))$ scalarly in $H\times H$, and by some calculations we get

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\big(R(y(t),z(t)),(y(t),z(t))\big)_{H\times H} \\
&\qquad + \frac{1}{2}\big(\|Ay(t)\|_H^2 + \|Az(t)\|_H^2 + \|B^*R(y(t),z(t))\|_{\mathbb{R}^{2N}}^2\big) \\
&\quad \le \|\mathcal{G}(y(t))\|_{H\times H}\|R(y(t),z(t))\|_{H\times H} \\
&\quad \le C_R\|G(y(t))\|_H\big(\|Ay(t)\|_H + \|Az(t)\|_H\big)
\end{aligned} \tag{13.84}$$

for a. e. $t \in (0,T)$, where we used (13.57). We recall (13.78) and (13.10) and compute the right-hand side:

$$\begin{aligned}
I &= C_R\sqrt{C_G}\big(\|A^{1/2}y(t)\|_H^3 + \|\varphi_\infty\|_\infty^2\|A^{1/2}y(t)\|_H^2\big)\big(\|Ay(t)\|_H + \|Az(t)\|_H\big) \\
&\le C\|Ay(t)\|_H^2\|A^{1/2}y(t)\|_H^2 + C\|A^{1/2}y(t)\|_H^4 + C\|A^{1/2}y(t)\|_H^2\|Az(t)\|_H^2 \\
&\quad + C\|\varphi_\infty\|_\infty^2\|Ay(t)\|_H^2\|A^{1/2}y(t)\|_H \\
&\quad + C\|\varphi_\infty\|_\infty^2\|Ay(t)\|_H^2\|A^{1/2}y(t)\|_H + C\|\varphi_\infty\|_\infty^2\|A^{1/2}y(t)\|_H\|Az(t)\|_H^2.
\end{aligned}$$

Since $(y,z) \in S_\infty$, we get $\|A^{1/2}y(t)\|_H \le r$. Therefore

$$I \le C_1\big(\|Ay(t)\|_H^2 + \|Az(t)\|_H^2\big)\big(r^2 + \|\varphi_\infty\|_\infty^2 r\big),$$

where $C_1$ depends on the problem parameters, $\Omega$, and $\|F(\varphi_\infty)\|_\infty$. Replacing in (13.84), we conclude that for a. e. $t$,

$$\begin{aligned}
&\frac{d}{dt}\big(R(y(t),z(t)),(y(t),z(t))\big)_{H\times H} + \|Ay\|_H^2 + \|Az\|_H^2 \\
&\qquad \le C_1\big(\|Ay\|_H^2 + \|Az\|_H^2\big)\big(r^2 + \|\varphi_\infty\|_\infty^2 r\big).
\end{aligned} \tag{13.85}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Denoting $\overline{C_2} := 1 - C_1(r^2 + \|\varphi_\infty\|_\infty^2 r) > 0$, we get

$$r \le r_2 := \frac{-\overline{C_2}\|\varphi_\infty\|_\infty^2 + \sqrt{\overline{C_2}^2\|\varphi_\infty\|_\infty^4 + 4\overline{C_2}}}{2\overline{C_2}}. \tag{13.86}$$

We fix $\rho$ by (13.82), where

$$r \le r_0 := \min\{r_1, r_2\}. \tag{13.87}$$

In conclusion, we obtain

$$\frac{d}{dt}\big(R(y(t), z(t)), (y(t), z(t))\big)_{H\times H} + C_2\big(\|Ay(t)\|_H^2 + \|Az(t)\|_H^2\big) \le 0 \tag{13.88}$$

for a. e. $t \in (0, \infty)$. Recalling (13.10) and (13.56), we deduce that

$$\frac{d}{dt}\big(R(y(t), z(t)), (y(t), z(t))\big)_{H\times H} + C_2 c_0\big(R(y(t), z(t)), (y(t), z(t))\big)_{H\times H} \le 0 \tag{13.89}$$

for a. e. $t \in (0, \infty)$. This implies

$$\big(R(y(t), z(t)), (y(t), z(t))\big)_{H\times H} \le e^{-2kt}\big(R(y_0, z_0), (y_0, z_0)\big)_{H\times H}, \tag{13.90}$$

where $k := \frac{C_2 c_0}{2}$, and recalling again (13.56), we infer that

$$c_1\|(y(t), z(t))\|_{D(A^{1/2})\times D(A^{1/2})}^2 \le c_2 e^{-2kt}\|(y_0, z_0)\|_{D(A^{1/2})\times D(A^{1/2})}^2, \quad \text{a. e. } t > 0,$$

which leads to (13.74). The constants $k$, $c_1$, $c_2$ in the relation before depend on the problem parameters and $\|F(\varphi_\infty)\|_\infty$, which reduces here to $\|\varphi_\infty\|_\infty$.

In conclusion, $\rho$ can be fixed by (13.82), (13.83), and (13.86), depending on the problem parameters and $\|\varphi_\infty\|_\infty$, so that the stationary solution is exponentially stabilized. The proof is ended. □

## 13.3 Stabilization of the viscous Cahn–Hilliard system with the logarithmic potential

In this section, we discuss the stabilization of system (13.22)–(13.24) in which $F$ is the logarithmic potential (13.7). In this case, we first need to treat the problem for a regularized potential under the assumption

$$\varphi_\infty \text{ is analytic in } \Omega, \quad |\varphi_\infty(x)| < 1 - \varepsilon \quad \text{for } x \in \overline{\Omega}. \tag{13.91}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Let us fix a constant $\varepsilon \in (0,1)$, and let $\chi_\varepsilon \in C^\infty(\mathbb{R})$ be a smooth function satisfying

$$\chi_\varepsilon(r) = 1 \quad \text{if } |r| \le 1-\varepsilon,$$
$$\chi_\varepsilon(r) \in (0,1) \quad \text{if } 1-\frac{\varepsilon}{2} < |r| < 1-\frac{\varepsilon}{2},$$
$$\chi_\varepsilon(r) = 0 \quad \text{if } |r| \ge 1-\frac{\varepsilon}{2}.$$

We introduce the regularized potential

$$F_\varepsilon(r) := \chi_\varepsilon(r)F(r)$$

of class $C_0^\infty(\mathbb{R})$. We will replace the singular function $F$ in (13.22) by a regular function $F_\varepsilon(r)$ and prove for this new function results similar to those presented in Section 13.2.

We mention that, due to (13.91), $F_\varepsilon$ and its derivatives computed at $\varphi_\infty$ coincide with the derivatives of $F$ at $\varphi_\infty$, and so we can omit for them the subscript $\varepsilon$ (that is why in (13.92), we can keep the notation $F''(\varphi_\infty)$). Moreover, $F_\varepsilon$ and its derivatives are continuous (and bounded) on $[-1+\frac{\varepsilon}{2}, 1-\frac{\varepsilon}{2}]$, and they are zero outside this interval. We also specify that the derivatives of $F$ involved in the next computations are continuous on $\{r; |r| \le 1-\frac{\varepsilon}{2}\}$. Let us denote $C_F'' = \|F''\|_{L^\infty(-1+\frac{\varepsilon}{2},1-\frac{\varepsilon}{2})}$, $C_F''' = \|F'''\|_{L^\infty(-1+\frac{\varepsilon}{2},1-\frac{\varepsilon}{2})}$, and $C_F = \max\{C_F'', C_F'''\}$. Due to the definition of $F_\varepsilon$, it follows that

$$\|F_\varepsilon''\|_{L^\infty(-1+\frac{\varepsilon}{2},1-\frac{\varepsilon}{2})} \le CC_F'' \le CC_F, \quad \|F_\varepsilon'''\|_{L^\infty(-1+\frac{\varepsilon}{2},1-\frac{\varepsilon}{2})} \le CC_F.$$

The nonlinear system (13.22)–(13.24) now becomes

$$y_t + \frac{\nu}{\tau^2}(A + A^{-1} - 2)y - \frac{1}{\tau}(A^{-1} - I)(F''(\varphi_\infty)y) + \frac{\gamma}{\tau}(A^{-1} - I)z - \frac{\gamma l}{\tau}(A^{-1} - I)y = f_\omega v + \frac{1}{\tau}(A^{-1} - I)F_{r,\varepsilon}(y) \quad \text{in } (0,\infty)\times\Omega, \tag{13.92}$$

$$z_t + \frac{1}{\tau}(A - I)z + \frac{l}{\tau}(I - A)y = f_\omega u \quad \text{in } (0,\infty)\times\Omega, \tag{13.93}$$

$$y(0) = y_0, \quad z(0) = z_0 \quad \text{in } \Omega, \tag{13.94}$$

where $F_{r,\varepsilon}$ is the second-order remainder in the Taylor expansion of $F_\varepsilon'(y+\varphi_\infty)$, written in the integral form

$$F_{r,\varepsilon}(y) = y^2 \int_0^1 (1-s)F_\varepsilon'''(\varphi_\infty + sy)ds. \tag{13.95}$$

Therefore we get the same linearized system (13.25)–(13.27) with the corresponding operator $\mathcal{A}$ given by (13.30), and, consequently, all results in Sections 13.2.1 and 13.2.2

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

remain valid, with the constants depending on the problem parameters, $\Omega$, and on $C_F$ via $\|F''(\varphi_\infty)\|_\infty$.

The nonlinear system in the closed-loop form now is

$$\begin{aligned} \frac{d}{dt}\big(y(t),z(t)\big) + \mathcal{A}\big(y(t),z(t)\big) &= \mathcal{G}_\varepsilon\big(y(t)\big) - BB^* R\big(y(t),z(t)\big), \quad \text{a. e. } t>0, \\ \big(y(0),z(0)\big) &= (y_0,z_0), \end{aligned} \tag{13.96}$$

where $\mathcal{G}_\varepsilon(y) = (G_\varepsilon(y),0)$, and $G_\varepsilon(y) = \frac{1}{\tau}(A^{-1}-I)F_{r,\varepsilon}(y)$. Following the same arguments as in the proofs presented in the previous section, but with some modified computations due to the current expression of the potential, we can state the following result.

**Theorem 13.3.1.** *Let $(y_0,z_0)\in D(A^{1/2})\times D(A^{1/2})$. There exists $\rho$ (depending on the problem parameters, $\Omega$, and $C_F$) such that if* (13.72) *takes place, then the closed-loop system* (13.96) *corresponding to $F_\varepsilon$ has a unique solution in the spaces* (13.73). *The solution is exponentially stable and satisfies* (13.74) *for some positive constants $k_\infty$ and $C_\infty$, which depend on the problem parameters, $\Omega$, and $C_F$.*

Theorem 13.3.1 is a general stabilization result for a function $F_\varepsilon$, which and its derivatives up to the third order are continuous. This implies the following consequence for the system with logarithmic potential $F$.

**Theorem 13.3.2.** *Let $\varepsilon\in(0,1)$ be arbitrary but fixed. For all pairs $(y_0,z_0)\in D(A^{1/2})\times D(A^{1/2})$ satisfying $\|y_0\|_{D(A^{1/2})}+\|z_0\|_{D(A^{1/2})}\le\rho$, the closed-loop system* (13.96) *corresponding to the logarithmic potential $F$ has in the one-dimensional case a unique solution belonging to the spaces* (13.73). *The solution is exponentially stable and satisfies* (13.74).

*Proof.* We recall the result of Theorem 13.3.1 for system (13.92)–(13.94) corresponding to $F_\varepsilon$ and write (13.92) in the form

$$\begin{aligned} &y_t + \frac{\nu}{\tau^2}(A+A^{-1}-2)y - \frac{1}{\tau}(A^{-1}-I)\big(F_\varepsilon'(y+\varphi_\infty)-F_\varepsilon'(\varphi_\infty)\big) \\ &\quad + \frac{\gamma}{\tau}(A^{-1}-I)z - \frac{\gamma l}{\tau}(A^{-1}-I)y = f_\omega v \quad \text{in } (0,\infty)\times\Omega, \end{aligned} \tag{13.97}$$

which is exactly that before expanding $F_\varepsilon'(y+\varphi_\infty)$ in Taylor series. Due to (13.91), in (13.97), we can write $F'(\varphi_\infty)$ instead of $F_\varepsilon'(\varphi_\infty)$. Theorem 13.3.1 claims that there exists $\rho$ given by (13.82) such that if the initial datum is in the ball with radius $\rho$, then

$$\big\|y(t)\big\|_{D(A^{1/2})} + \big\|z(t)\big\|_{D(A^{1/2})} \le C_\infty e^{-k_\infty t}\rho. \tag{13.98}$$

For $d=1$, we have that $H^1(\Omega)$ is compact in $C(\overline{\Omega})$ and

$$\big|y(t)\big| \le \big\|y(t)\big\|_{C(\overline{\Omega})} \le C_\Omega \big\|y(t)\big\|_{D(A^{1/2})} \le C_\Omega C_\infty e^{-k_\infty t}\rho, \tag{13.99}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

and so (13.98) implies that $|y(t)| \to 0$ as $t \to \infty$. For a sufficient large $t > \frac{1}{k_\infty} \ln \frac{\rho C_\infty C_\Omega}{1-\varepsilon}$, we can find the values of $|y(t)|$ in a ball with radius less than $1-\varepsilon$. By a suitable choice of a new $\rho$ the solution remains less than $1-\varepsilon$ for all $t > 0$. Because $|\varphi_\infty| < 1-\varepsilon$, we can write $|\varphi_\infty| \le 1-\varepsilon-\delta$ with $\delta \in (0, 1-\varepsilon)$, and so we can impose in (13.99) that

$$|y(t)| \le C_\Omega C_\infty e^{-k_\infty t}\rho \le C_\Omega C_\infty \rho \le \delta \quad \text{for all } t \ge 0.$$

This happens if $\rho \le \frac{\delta}{C_\Omega C_\infty}$, and recalling (13.82), we can set new $\rho$ such that

$$\rho \le \min\left\{\frac{\delta}{C_\Omega C_\infty}, r\sqrt{\frac{c_1}{2c_2}}\right\}$$

with $r < r_0$ by (13.87). Then we have

$$|y(t) + \varphi_\infty| < \delta + 1 - \varepsilon - \delta = 1 - \varepsilon \quad \text{for all } t \ge 0,$$

and, consequently, we can write $F_\varepsilon'(y + \varphi_\infty) = F'(y + \varphi_\infty)$ in (13.97). In conclusion, our solution $y(t)$ in fact satisfies system (13.97), (13.93)–(13.94) corresponding to the function $F$, and we have the stabilization result in the one-dimensional case. □

## 13.4 Stabilization of the nonviscous Cahn–Hilliard system with a regular potential

Finally, we discuss the feedback stabilization for system (13.1)–(13.5) with $\tau = 0$. In this case the proposed controller is $(f_\omega v, f_\omega u)$ with the same $f_\omega$ defined in (13.16). We observe that a suitable function transformation $\sigma := \alpha_0(\theta + l\varphi)$ with $\alpha_0 = \sqrt{\frac{\gamma}{l}}$ chosen such that $\frac{\gamma}{\alpha_0} = \alpha_0 l =: \gamma_0 > 0$ will give the possibility to work later on with a self-adjoint operator acting on the linear part of the system.

Writing system (13.1)–(13.5) in the variables $\varphi$ and $\sigma$ and using the notation $l_0 := \gamma l$, we get the equivalent nonlinear system

$$\varphi_t + \nu\Delta^2\varphi - \Delta F'(\varphi) - l_0\Delta\varphi + \gamma_0\Delta\sigma = f_\omega v \quad \text{in } (0,\infty)\times\Omega, \tag{13.100}$$

$$\sigma_t - \Delta\sigma + \gamma_0\Delta\varphi = f_\omega u \quad \text{in } (0,\infty)\times\Omega, \tag{13.101}$$

$$\frac{\partial\varphi}{\partial n} = \frac{\partial\Delta\varphi}{\partial n} = \frac{\partial\sigma}{\partial n} = 0 \quad \text{in } (0,\infty)\times\partial\Omega, \tag{13.102}$$

$$\varphi(0) = \varphi_0, \quad \sigma(0) = \sigma_0 := \alpha_0(\theta_0 + l\varphi_0) \quad \text{in } \Omega. \tag{13.103}$$

We recall that a stationary solution to the uncontrolled system has the properties that $\theta_\infty$ is constant and $\varphi_\infty \in H^4(\Omega) \subset C^2(\overline{\Omega})$. We develop $F'(y + \varphi_\infty)$ in Taylor expan-

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

sion and rewrite (13.100) as

$$y_t + \nu\Delta^2 y - \Delta(F''(\varphi_\infty)y) - l_0\Delta y + \gamma_0\Delta z = \Delta F_r(y) + f_\omega v, \tag{13.104}$$

where $F_r(y)$ is the second-order remainder. Then we define

$$\overline{F''_\infty} = \frac{1}{m_\Omega}\int_\Omega F''(\varphi_\infty(\xi))d\xi = \frac{3}{m_\Omega}\int_\Omega \varphi^2_\infty(\xi)d\xi - 1, \tag{13.105}$$

where $m_\Omega$ is the measure of $\Omega$. Thus we have $F''(\varphi_\infty(x)) = \overline{F''_\infty} + g(x)$, where

$$g(x) := \frac{1}{m_\Omega}\int_\Omega (F''(\varphi_\infty(x)) - F''(\varphi_\infty(\xi)))d\xi = \frac{3}{m_\Omega}\int_\Omega (\varphi^2_\infty(x) - \varphi^2_\infty(\xi))d\xi. \tag{13.106}$$

Plugging (13.105) into (13.104), we get the following equivalent form of the nonlinear system (13.100)–(13.103):

$$y_t + \nu\Delta^2 y - F_l\Delta y + \gamma_0\Delta z = \Delta(F_r(y) + g(x)y) + f_\omega v \quad \text{in } (0,\infty)\times\Omega, \tag{13.107}$$

$$z_t - \Delta z + \gamma_0\Delta y = f_\omega u \quad \text{in } (0,\infty)\times\Omega, \tag{13.108}$$

$$y(0) = y_0, \quad z(0) = z_0 \quad \text{in } \Omega, \tag{13.109}$$

$$\frac{\partial y}{\partial n} = \frac{\partial \Delta y}{\partial n} = \frac{\partial z}{\partial n} = 0 \quad \text{in } (0,\infty)\times\partial\Omega, \tag{13.110}$$

where $F_l = \overline{F''_\infty} + l_0$. We note that $F_l$ also depends on $\Omega$ and $\|\varphi_\infty\|_{L^2(\Omega)}$.

Now we extract the linear system

$$y_t + \nu\Delta^2 y - F_l\Delta y + \gamma_0\Delta z = f_\omega v \quad \text{in } (0,\infty)\times\Omega, \tag{13.111}$$

$$z_t - \Delta z + \gamma_0\Delta y = f_\omega u \quad \text{in } (0,\infty)\times\Omega, \tag{13.112}$$

$$y(0) = y_0, \quad z(0) = z_0 \quad \text{in } \Omega, \tag{13.113}$$

$$\frac{\partial y}{\partial n} = \frac{\partial \Delta y}{\partial n} = \frac{\partial z}{\partial n} = 0 \quad \text{in } (0,\infty)\times\partial\Omega. \tag{13.114}$$

The associated operator $\mathcal{A} : D(\mathcal{A}) \subset \mathcal{H} \to \mathcal{H}$ defined by

$$\mathcal{A} = \begin{bmatrix} \nu\Delta^2 - F_l\Delta & \gamma\Delta \\ \gamma\Delta & -\Delta \end{bmatrix}$$

has the domain

$$D(\mathcal{A}) = \Big\{ w = (y,z) \in H^2(\Omega)\times H^1(\Omega);\ \mathcal{A}w \in \mathcal{H},$$
$$\frac{\partial y}{\partial n} = \frac{\partial \Delta y}{\partial n} = \frac{\partial z}{\partial n} = 0 \text{ on } \partial\Omega \Big\}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

and is self-adjoint. We write the linearized system in the abstract form

$$\frac{d}{dt}\big(y(t),z(t)\big)+\mathcal{A}\big(y(t),z(t)\big)=f_\omega U(t),\quad \text{a. e. } t\in(0,\infty),\qquad (13.115)$$
$$\big(y(0),z(0)\big)=(y_0,z_0),$$

where $U(t)=(v(t),u(t))$.

Proposition 13.2.1 follows exactly as in Section 13.2. In this case, the eigenvalues are real and semisimple, that is, $\mathcal{A}$ is diagonalizable (see [14], p. 59). The eigenvectors corresponding to distinct eigenvalues are orthogonal. Then, orthogonalizing the system $\{(\varphi_i,\psi_i)\}_i$ in the space $\mathcal{H}$, we may assume that it is orthonormal and complete in $\mathcal{H}$. Moreover, since the resolvent of $\mathcal{A}$ is compact, there exists a finite number of nonpositive eigenvalues (see [14], p. 187), and every eigenvalue is repeated according to its multiplicity. Let $N$ be the number of these nonpositive eigenvalues, that is, $\lambda_i\le 0$ for $i=1,\dots,N$. This will simplify the proof of Proposition 13.2.2, which we now prove in a real space.

The representation of the feedback controller as $f_\omega U(t)=-BB^*R(y(t),z(t))$ and the characterization of its properties are proved like in Propositions 13.2.3 and 13.2.4, with some appropriate modifications due to the actual form of the operator.

The nonlinear closed-loop system reads

$$\frac{d}{dt}\big(y(t),z(t)\big)+\mathcal{A}\big(y(t),z(t)\big)=\mathcal{G}\big(y(t)\big)-BB^*R\big(y(t),z(t)\big),\quad \text{a. e. } t>0,$$
$$\big(y(0),z(0)\big)=(y_0,z_0),\qquad (13.116)$$

where $\mathcal{G}(y(t))=(G(y(t)),0)$, and

$$G(y)=\Delta F_r(y)+\Delta\big(g(x)y\big).\qquad (13.117)$$

A visible difference occurs in the proof of the corresponding Theorem 13.2.5, because the determination of $r_1$, $r_2$, and $\rho$ implies come conditions induced by the estimates for the term $\Delta(g(x)y)$ in (13.117).

All these lead to the statement of the feedback stabilization result in the case of nonviscous Cahn–Hilliard system in the following:

**Theorem 13.4.1.** *We set* $\chi_\infty:=\|\nabla\varphi_\infty\|_\infty+\|\Delta\varphi_\infty\|_\infty$. *Then there exists* $\chi_0>0$ *(depending on the problem parameters, the domain, and* $\|\varphi_\infty\|_\infty$*) such that the following holds. If* $\chi_\infty\le\chi_0$, *then there exists* $\rho$ *such that for all pairs* $(y_0,z_0)\in D(A^{1/2})\times D(A^{1/4})$ *with* $\|y_0\|_{D(A^{1/2})}+\|z_0\|_{D(A^{1/4})}\le\rho$, *the closed-loop system* (13.116) *has a unique solution*

$$(y,z)\in C\big([0,\infty);H\times H\big)\cap L^2\big(0,\infty;D(A^{3/2})\times D(A^{3/4})\big)$$
$$\cap\, W^{1,2}\big(0,\infty;(D(A^{1/2})\times D(A^{1/4}))'\big),$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*which is exponentially stable, that is,*

$$\|y(t)\|_{D(A^{1/2})} + \|z(t)\|_{D(A^{1/4})} \leq C_P e^{-kt}(\|y_0\|_{D(A^{1/2})} + \|z_0\|_{D(A^{1/4})})$$

*for some positive constants k and* $C_P$.

In the previous relations, the positive constants $k$ and $C_P$ depend on $\Omega$, the problem parameters, and $\|\varphi_\infty\|_\infty$. In addition, $C_P$ depends on the full norm $\|\varphi_\infty\|_{2,\infty}$.

A similar result of feedback stabilization can be proved for the singular logarithmic potential in the one-dimensional case, as discussed in [17], except for the value $\chi_\infty$, which in this case is $\chi_\infty := \|\nabla\varphi_\infty\|_\infty + \|\nabla\varphi_\infty\|_\infty^2 + \|\Delta\varphi_\infty\|_\infty$.

# Bibliography


[1] V. Barbu. *Mathematical Methods in Optimization of Differential Systems*. Kluwer Academic Publishers, Dordrecht, 1994.

[2] V. Barbu. *Nonlinear Differential Equations of Monotone Type in Banach Spaces*. Springer, London, New York, 2010.

[3] V. Barbu. *Stabilization of Navier–Stokes Flows*. Springer-Verlag, London, 2011.

[4] V. Barbu, P. Colli, G. Gilardi, and G. Marinoschi. Feedback stabilization of the Cahn–Hilliard type system for phase separation. *J. Differ. Equ.*, 262:2286–2334, 2017.

[5] V. Barbu, I. Lasiecka, and R. Triggiani. Tangential boundary stabilization of Navier–Stokes equations. *Mem. Am. Math. Soc.*, 181:852, 2006.

[6] V. Barbu and R. Triggiani. Internal stabilization of Navier–Stokes equations with finite-dimensional controllers. *Indiana Univ. Math. J.*, 53:1443–1494, 2004.

[7] V. Barbu and G. Wang. Internal stabilization of semilinear parabolic systems. *J. Math. Anal. Appl.*, 285:387–407, 2003.

[8] T. Breiten, K. Kunisch, and L. Pfeiffer. Feedback stabilization of the two-dimensional Navier–Stokes equations by value function approximation. *Appl. Math. Optim.*, 80:599–641, 2019.

[9] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, New York, Dordrecht, Heidelberg, London, 2011.

[10] G. Caginalp. An analysis of a phase field model of a free boundary. *Arch. Ration. Mech. Anal.*, 92:205–245, 1986.

[11] J. W. Cahn and J. E. Hilliard. Free energy of a nonuniform system I. Interfacial free energy. *J. Chem. Phys.*, 2:258–267, 1958.

[12] P. Colli, G. Gilardi and G. Marinoschi. A boundary control problem for a possibly singular phase field system with dynamic boundary conditions. *J. Math. Anal. Appl.*, 434:432–463, 2016.

[13] D. Fujiwara. Concrete characterization of the domains of fractional powers of some elliptic differential operators of the second order. *Proc. Jpn. Acad.*, 43:82–86, 1967.

[14] T. Kato. *Perturbation Theory for Linear Operators*. Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1984.

[15] J. Kim, S. Lee, Y. Choi, S.-M. Lee, and D. Jeong. Basic principles and practical applications of the Cahn–Hilliard equation. *Math. Probl. Eng.*, 2016:Article ID 9532608, 11, 2016.

[16] J. L. Lions. *Optimal Control of Systems Governed by Partial Differential Equations*, volume 170 of Die Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, 1971.


EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[17] G. Marinoschi. A note on the feedback stabilization of a Cahn–Hilliard type system with a singular logarithmic potential. In A. Favini, P. Colli, E. Rocca, G. Schimperna, and J. Sprekels, editors, *Solvability, Regularity, Optimal Control of Boundary Value Problems for PDEs*, pages 357–377, volume 22 of Springer INdAM Series. Springer, 2017.

[18] G. Marinoschi. Internal feedback stabilization of a Cahn–Hilliard system with viscosity effects. *Pure Appl. Funct. Anal.*, 3(1):107–135, 2018.

[19] A. Pazy. *Semigroups of Linear Operators and Applications to Partial Differential Equations*. Springer Verlag, New York, 1983.

[20] J.-P. Raymond and L. Thevenet. Boundary feedback stabilization of the two dimensional Navier–Stokes equations with finite dimensional controllers. *Discrete Contin. Dyn. Syst.*, 27:1159–1187, 2010.

[21] R. Triggiani. Boundary feedback stabilizability of parabolic equations. *Appl. Math. Optim.*, 6:201–220, 1980.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Philipp A. Guth, Claudia Schillings, and Simon Weissmann

# 14 Ensemble Kalman filter for neural network-based one-shot inversion

**Abstract:** We study the use of novel techniques arising in machine learning for inverse problems. Our approach replaces the complex forward model by a neural network, which is trained simultaneously in a one-shot sense when estimating the unknown parameters from data, i. e., the neural network is trained only for the unknown parameter. By establishing a link to the Bayesian approach to inverse problems we develop an algorithmic framework that ensures the feasibility of the parameter estimate with respect to the forward model. We propose an efficient, derivative-free optimization method based on variants of the ensemble Kalman inversion. Numerical experiments show that the ensemble Kalman filter for neural network-based one-shot inversion is a promising direction combining optimization and machine learning techniques for inverse problems.

**Keywords:** inverse problems, ensemble Kalman inversion, neural networks, one-shot optimization

**MSC 2010:** 65N21, 62F15, 65N75, 90C56

## 14.1 Introduction

Inverse problems arise in almost all areas of applications, e. g., biological problems, engineering, and environmental systems. The integration of data can substantially reduce the uncertainty in predictions based on the model and is therefore indispensable in many applications. Advances in machine learning provide exciting potential to complement and enhance simulators for complex phenomena in the inverse setting. We propose a novel approach to inverse problems by approximating the forward problem with neural networks, i. e., the computationally intense forward problem will be replaced by a neural network. However, the neural network when used as a surrogate model needs to be trained beforehand to guarantee good approximations of the forward problem for arbitrary inputs of the unknown coefficient. To reduce the costs

**Acknowledgement:** PG and SW are grateful to the DFG RTG1953 Statistical Modeling of Complex Systems and Processes for funding their research.

**Philipp A. Guth, Claudia Schillings,** University of Mannheim, School of Business Informatics and Mathematics, B6, 28-29, 68159 Mannheim, Germany, e-mails: pguth@uni-mannheim.de, c.schillings@uni-mannheim.de
**Simon Weissmann,** University of Heidelberg, Interdisciplinary Center for Scientific Computing, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany, e-mail: simon.weissmann@uni-heidelberg.de

https://doi.org/10.1515/9783110695984-014

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

associated with building the surrogate in the entire parameter space, we suggest to solve the inverse problem and train the neural network simultaneously in a one-shot fashion. From a computational point of view, this approach has the potential to reduce the overall costs significantly, as the neural network serves as a surrogate of the forward model only in the unknown parameter estimated from the data and not in the entire parameter space. The ensemble Kalman inversion (EKI) as a derivative-free optimizer is applied to the resulting optimization problem.

### 14.1.1 Background and literature overview

The goal of computation is to solve the following inverse problem: Recover the unknown parameter $u \in \mathcal{X}$ in an abstract model

$$M(u, p) = 0 \tag{14.1}$$

from a finite number of observation of the state $p \in \mathcal{V}$ given by

$$O(p) = y \in \mathbb{R}^{n_y}, \tag{14.2}$$

which might be subject to noise. We denote by $\mathcal{X}$, $\mathcal{V}$, and $\mathcal{W}$ Banach spaces. The operator $M : \mathcal{X} \times \mathcal{V} \to \mathcal{W}$ describes the underlying model, typically a PDE or ODE in the applications of interest. The variable $p$ denotes the state of the model and is defined on a domain $D \subset \mathbb{R}^d$, $d \in \mathbb{N}$. The operator $O : \mathcal{V} \to \mathbb{R}^{n_y}$ is the observation operator, mapping the state variables $p$ to the observations. Classical methods to inverse problems are based on an optimization approach of the (regularized) data misfit; see, e. g., [12, 25]. One-shot (all-at-once) approaches are well established in the context of PDE-constrained optimization (see, e. g., [5]) and have been recently also introduced in the inverse setting [23, 24]. The idea is to solve the underlying model equation simultaneously with the optimality conditions. This is in contrast to the so-called reduced or black-box methods, which formulate the minimization as an unconstrained optimization problem via the solution operator of (14.1). The connection of the optimization approach to the Bayesian setting via the maximum a posteriori estimate is well established in the finite-dimensional setting [22] as well as in the infinite-dimensional setting for certain prior classes; see [1, 8, 9, 18]. We refer to [22, 38] for more detail on the Bayesian approach to inverse problems. Recently, data-driven approaches have been introduced to the inverse setting to reduce the computational complexity in case of highly complex forward models and to improve models in case of limited understanding of the underlying process [2]. Neural networks in the inverse framework have been successfully applied in the case of parametric holomorphic forward models [19] and in the case of limited knowledge of the underlying model [30–32, 37, 39]. To train neural networks, gradient-based methods are typically used [16]. The ensemble Kalman inversion (EKI) (see, e. g., [21, 33, 34]) has been recently applied as gradient-free optimizer for the training of neural networks in [17, 27].

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

### 14.1.2 Our contribution

The goal of this work is twofold:

- We formulate the inverse problem in a one-shot fashion and establish the connection to the Bayesian setting. In particular, the Bayesian viewpoint allows us to incorporate model uncertainty and gives a natural way to define the regularization parameters. In case of an exact forward model the vanishing noise can be interpreted as a penalty method. This allows us to establish convergence results of the one-shot formulation as an unconstrained optimization problem to the corresponding (regularized) solution of the reduced optimization problem (with exact forward model). The (numerical approximation of the) forward problem is replaced by a neural network in the one-shot formulation, i. e., the neural network does not have to be trained in advance for all potential parameters.
- Secondly, we show that the EKI can be effectively used to solve the resulting optimization problem. We provide a convergence analysis in the linear setting. To enhance the performance, we modify the algorithm motivated by the continuous version of the EKI. Numerical experiments demonstrate the robustness (also in the nonlinear setting) of the proposed algorithm.

### 14.1.3 Outline

In Section 14.2, we discuss the optimization and Bayesian approach to inverse problems, introduce the one-shot formulation, and establish the connection to vanishing noise and penalty methods in case of exact forward models. In Section 14.3, we give an overview of neural networks used to approximate the forward problem. In Section 14.4, we discuss and analyze the EKI as a derivative-free optimizer for the one-shot formulation. Finally, in Section 14.5, we present numerical experiments demonstrating the effectiveness of the proposed approach. We conclude in Section 14.6 with an outlook to future work.

### 14.1.4 Notation

We denote by $\|\cdot\|$ the Euclidean norm and by $\langle\cdot,\cdot\rangle$ the corresponding inner product. For a given symmetric positive definite matrix (s. p. d.) $A$, the weighted norm $\|\cdot\|_A$ is defined by $\|\cdot\|_A = \|A^{-1/2}\cdot\|$, and the weighted inner product by $\langle\cdot,\cdot\rangle_A = \langle\cdot, A^{-1}\cdot\rangle$.

## 14.2 Problem formulation

We introduce in the following the optimization and Bayesian approach to an abstract inverse problem of the form (14.1)–(14.2).

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Throughout this paper, we derive the methods and theoretical results under the assumption that $\mathcal{X}$, $\mathcal{W}$, and $\mathcal{V}$ are finite-dimensional, i. e., we assume that the forward problem $M(u,p) = 0$ is discretized by a suitable numerical scheme and the parameter space is finite-dimensional, possibly after dimension truncation. Though most of the ideas and results can be generalized to the infinite-dimensional setting, we avoid the technicalities arising from the infinite-dimensional setting and focus on the discretized problem, i. e., $\mathcal{X} = \mathbb{R}^{n_u}$, $\mathcal{V} = \mathbb{R}^{n_p}$, and $\mathcal{W} = \mathbb{R}^{n_w}$.

### 14.2.1 Optimization approach to the inverse problem

The optimization approach leads to the following problem:

$$\min_{u,p} \|O(p) - y\|^2_{\Gamma_{\text{obs}}} \tag{14.3}$$

$$\text{s. t. } M(u,p) = 0, \tag{14.4}$$

i. e., we minimize the data misfit in a suitable weighted norm with s. p. d. $\Gamma_{\text{obs}} \in \mathbb{R}^{n_y \times n_y}$, given that the forward model is satisfied. Due to the ill-conditioning of the inverse problem, a regularization term on the unknown parameters is often introduced to stabilize the optimization, i. e., we consider

$$\min_{u,p} \|O(p) - y\|^2_{\Gamma_{\text{obs}}} + \alpha_1 \mathcal{R}_1(u) \tag{14.5}$$

$$\text{s. t. } M(u,p) = 0, \tag{14.6}$$

where the regularization is denoted by $\mathcal{R}_1 : \mathcal{X} \to \mathbb{R}$, and the positive scalar $\alpha_1 > 0$ is usually chosen according to prior knowledge on the unknown parameter $u$. We will comment on the motivation of the regularization via the Bayesian approach in the following section. To do so, we first introduce the so-called reduced problems of (14.3)–(14.4) and (14.5)–(14.6). The forward model $M(u,p) = 0$ is typically a well-posed problem, in the sense that for each parameter $u \in \mathcal{X}$, there exists a unique state $p \in \mathcal{V}$ such that $M(u,p) = 0$ in $\mathcal{W}$. Introducing the solution operator $S : \mathcal{X} \to \mathcal{V}$ such that $M(u, S(u)) = 0$, we can reformulate the optimization problems (14.3)–(14.4) and (14.5)–(14.6) as an unconstrained optimization problem

$$\min_{u \in \mathcal{X}} \|O(S(u)) - y\|^2_{\Gamma_{\text{obs}}} \tag{14.7}$$

and

$$\min_{u \in \mathcal{X}} \|O(S(u)) - y\|^2_{\Gamma_{\text{obs}}} + \alpha_1 \mathcal{R}_1(u),$$

respectively.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

### 14.2.2 Bayesian approach to the inverse problem

Adopting the Bayesian approach to inverse problems, we view the unknown parameters $u$ as an $\mathcal{X}$-valued random variable with prior distribution $\mu_0$. The noise in the observations is assumed to be additive and described by a random variable $\eta \sim \mathcal{N}(0, \Gamma_{\text{obs}})$ with s. p. d. $\Gamma_{\text{obs}} \in \mathbb{R}^{n_y \times n_y}$, i. e.,

$$y = O(S(u)) + \eta.$$

Further, we assume that the noise $\eta$ is stochastically independent of $u$. By Bayes' theorem we obtain the posterior distribution

$$\mu^*(\mathrm{d}u) \propto \exp\left(-\frac{1}{2}\left\|O(S(u)) - y\right\|^2_{\Gamma_{\text{obs}}}\right)\mu_0(\mathrm{d}u),$$

the conditional distribution of the unknown given the observation $y$.

### 14.2.3 Connection between the optimization and the Bayesian approach

The optimization approach leads to a point estimate of the unknown parameters, whereas the Bayesian approach computes the conditional distribution of the unknown parameters given the data, the posterior distribution, as a solution of the inverse problem. Since the computation or approximation of the posterior distribution is prohibitively expensive in many applications, one often restores to point estimates such as the maximum a posteriori (MAP) estimate, the most likely point of the unknown parameters. Denoting by $\rho_0$ the Lebesgue density of the prior distribution, the MAP estimate is defined as

$$\arg\max_{u \in \mathcal{X}} \exp\left(-\frac{1}{2}\left\|O(S(u)) - y\right\|^2_{\Gamma_{\text{obs}}}\right)\rho_0(u).$$

Assuming a Gaussian prior distribution, i. e., $\mu_0 = \mathcal{N}(u_0, C)$, the MAP estimate is given by the solution of the following minimization problem:

$$\min_u \ \frac{1}{2}\left\|O(S(u)) - y\right\|^2_{\Gamma_{\text{obs}}} + \frac{1}{2}\|u - u_0\|^2_C.$$

The Gaussian prior assumption leads to a Tikhonov-type regularization in the objective function, whereas the first term in the objective function results from the Gaussian assumption on the noise. We refer the interested reader to [10, 22] for more detail on the MAP estimate.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

### 14.2.4 One-shot formulation for inverse problems

Black-box methods are based on the reduced formulation (14.7) assuming that the forward problem can be solved exactly in each iteration. Here we follow a different approach, which simultaneously solves the forward and optimization problem. Various names for the simultaneous solution of the design and state equation exist: one-shot method, all-at-once, piggy-back iterations, etc.. We refer the reader to [5] and the references therein for more detail.

Following the one-shot ideas, we seek to solve the problem

$$F(u,p) = \begin{pmatrix} M(u,p) \\ O(p) \end{pmatrix} = \begin{pmatrix} 0 \\ y \end{pmatrix} =: \tilde{y}.$$

Due to the noise in the observations, we rather consider

$$y = O(p) + \eta_{\text{obs}}$$

with normally distributed noise $\eta_{\text{obs}} \sim \mathcal{N}(0, \Gamma_{\text{obs}})$, s. p. d. $\Gamma_{\text{obs}} \in \mathbb{R}^{n_y \times n_y}$. Similarly, we assume that

$$0 = M(u,p) + \eta_{\text{model}},$$

i. e., we assume that the model error can be described by $\eta_{\text{model}} \sim \mathcal{N}(0, \Gamma_{\text{model}})$, s. p. d. $\Gamma_{\text{model}} \in \mathbb{R}^{n_w \times n_w}$. Thus we obtain the problem

$$\tilde{y} = F(u,p) + \begin{pmatrix} \eta_{\text{model}} \\ \eta_{\text{obs}} \end{pmatrix}.$$

The MAP estimate is then computed by the solution of the following minimization problem:

$$\min_{u,p} \frac{1}{2} \left\| F(u,p) - \tilde{y} \right\|_{\Gamma}^2 + \alpha_1 \mathcal{R}_1(u) + \alpha_2 \mathcal{R}_2(p),$$

where $\mathcal{R}_1 : \mathcal{X} \to \mathbb{R}$ and $\mathcal{R}_2 : \mathcal{V} \to \mathbb{R}$ are regularizations of the parameter $u \in \mathcal{X}$ and the state $p \in \mathcal{V}$, $\alpha_1, \alpha_2 > 0$, and $\Gamma = \left( \begin{smallmatrix} \Gamma_{\text{model}} & 0 \\ 0 & \Gamma_{\text{obs}} \end{smallmatrix} \right) \in \mathbb{R}^{(n_w+n_y)\times(n_w+n_y)}$.

**Remark 14.2.1.** Note that the proposed approach does not rely on a Gaussian noise model for the forward problem, i. e., non-Gaussian models can be straightforwardly incorporated. Then the Bayesian viewpoint may guide the choice of the regularization parameter (or function) in the optimization approach via the MAP estimate. The model error is typically estimated from experimental data or more complex models; see [20, 26]. We focus here on the Gaussian setting, as the one-shot approach for inverse problems is typically formulated in a least-squares fashion (in particular, when using neural networks as surrogate models of the forward problem [30, 31]). The focus

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

of this work will be on the development of a methodology, which allows the forward problem to satisfy exactly. This will be achieved by establishing the connection to the Bayesian setting and working in the vanishing noise setting.

### 14.2.5 Vanishing noise and penalty methods

In case of an exact forward model, i. e., in case the forward equation is supposed to be satisfied exactly with $M(u,p) = 0$, this can be modeled in the Bayesian setting by vanishing noise. To illustrate this idea, we consider a parameterized noise covariance model $\Gamma_{\text{model}} = \gamma\hat{\Gamma}_{\text{model}}$ for $\gamma \in \mathbb{R}_+$ and given the s. p. d. matrix $\hat{\Gamma}_{\text{model}}$. The limit as $\gamma \to 0$ corresponds to the vanishing noise setting and can be interpreted as reducing the uncertainty in our model. The MAP estimate in the one-shot framework is thus given by

$$\min_{u,p} \frac{1}{2}\|O(p) - y\|^2_{\Gamma_{\text{obs}}} + \frac{\lambda}{2}\|M(u,p)\|^2_{\hat{\Gamma}_{\text{model}}} + \alpha_1\mathcal{R}_1(u) + \alpha_2\mathcal{R}_2(p) \tag{14.8}$$

with $\lambda = 1/\gamma$. This form reveals a close connection to penalty methods, which attempt to solve constrained optimization problems such as (14.3)–(14.4) by sequentially solving unconstrained optimization problems of the form (14.8) for an increasing sequence of penalty parameters $\lambda$. The following well-known result on the convergence of the resulting algorithm can be found, e. g., in [3].

**Proposition 14.2.1.** *Let the observation operator $O$, the forward model $M$, and the regularization functions $\mathcal{R}_1$, $\mathcal{R}_2$ be continuous, and let the feasible set $\{(u,p) \mid M(u,p) = 0\}$ be nonempty. For $k = 0, 1, \ldots$, let $(u_k, p_k)$ denote a global minimizer of*

$$\min_{u,p} \frac{1}{2}\|O(p) - y\|^2_{\Gamma_{\text{obs}}} + \frac{\lambda_k}{2}\|M(u,p)\|^2_{\hat{\Gamma}_{\text{model}}} + \alpha_1\mathcal{R}_1(u) + \alpha_2\mathcal{R}_2(p)$$

*with $(\lambda_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$ strictly increasing and $\lambda_k \to \infty$ as $k \to \infty$. Then every accumulation point of the sequence $(u_k, p_k)$ is a global minimizer of*

$$\begin{aligned} &\min_{u,p} \frac{1}{2}\|O(p) - y\|^2_{\Gamma_{\text{obs}}} + \alpha_1\mathcal{R}_1(u) + \alpha_2\mathcal{R}_2(p) \\ &\quad s.t.\ M(u,p) = 0. \end{aligned}$$

This classic convergence result ensures the feasibility of the estimates, i. e., the proposed approach is able to incorporate and exactly satisfy physical constraints in the limit. We mention also the possibility to consider exact penalty terms in the objective, corresponding to different noise models in the Bayesian setting. This will be subject to future work.

This setting will be the starting point of the incorporation of neural networks into the problem. Instead of minimizing with respect to the state $p$, we will approximate

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

the solution of the forward problem $p$ by a neural network $p_\theta$, where $\theta$ denotes the parameters of the neural network to be learned within this framework. Thus we obtain the corresponding minimization problem

$$\min_{u,\theta} \frac{1}{2}\left\|F(u,p_\theta)-\tilde{y}\right\|_\Gamma^2 + \alpha_1\mathcal{R}_1(u) + \alpha_2\mathcal{R}_2(p_\theta,\theta),$$

where $p_\theta$ denotes the state approximated by the neural network.

## 14.3 Neural networks

Following, e. g., [11, 29], we distinguish between the neural network architecture or neural network parameters as a set of weights and biases and the neural network, which is the associated mapping when an activation function is applied to the affine linear transformations defined by the neural network parameters:

The neural network architecture or the neural network parameters $\theta$ with input dimension $d \in \mathbb{N}$ and $L \in \mathbb{N}$ layers is a sequence of matrix–vector tuples

$$\theta = \big((W_\ell, b_\ell)\big)_{\ell=1}^L = \big((W_1,b_1),(W_2,b_2),\dots,(W_L,b_L)\big) \in \Theta,$$

where $\Theta := \times_{\ell=1}^L (\mathbb{R}^{N_\ell \times N_{\ell-1}} \times \mathbb{R}^{N_\ell}) \cong \mathbb{R}^{n_\theta}$ with $n_\theta = \sum_{\ell=1}^L (N_{\ell-1}+1)\,N_\ell$. Hence the number of neurons in layer $\ell$ is given by $N_\ell \in \mathbb{N}$ and $N_0 := d$, i. e., $W_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and $b_\ell \in \mathbb{R}^{N_\ell}$ for $\ell = 1,\dots,L$. Given an activation function $\varrho : \mathbb{R} \to \mathbb{R}$, we call the mapping

$$\begin{aligned} p_\theta^\varrho : \mathbb{R}^d &\to \mathbb{R}^{N_L} \\ x &\mapsto p_\theta^\varrho(x) := x_L \end{aligned}$$

defined by the recursion

$$\begin{aligned} x_0 &:= x, \\ x_\ell &:= \varrho(W_\ell x_{\ell-1} + b_\ell) \quad \text{for } \ell = 1,\dots,L-1, \\ x_L &:= W_L x_{L-1} + b_L \end{aligned}$$

a neural network $p_\theta^\varrho$ with activation function $\varrho$, where $\varrho$ is understood to act componentwise on vector-valued inputs, i. e., we mean $\varrho(z) := (\varrho(z_1),\dots,\varrho(z_n))$ for $z \in \mathbb{R}^n$. In the following the activation function is always the sigmoid function $\varrho = \frac{1}{1+e^{-x}}$. We will therefore waive the superscript indicating the activation function, i. e., we abbreviate $p_\theta^\varrho = p_\theta$.

This class of neural networks is often called feed-forward deep neural networks (DNNs) and has recently found success in forward [36] as well as Bayesian inverse [19] problems in the field of uncertainty quantification.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Based on the approximation results of polynomials by feed-forward DNNs [40], the authors in [36] derive bounds on the expression rate for multivariate real-valued functions depending holomorphically on a sequence $z = (z_j)_{j\in\mathbb{N}}$ of parameters. More specifically, the authors consider functions that admit sparse Taylor generalized polynomial chaos (g. p. c.) expansions, i. e., $s$-summable Taylor g. p. c. coefficients. Such functions arise as response surfaces of parametric PDEs, or in a more general setting from parametric operator equations; see, e. g., [35] and the references therein. Their main results is that these functions can be expressed with arbitrary accuracy $\delta > 0$ (uniform with respect to $z$) by DNNs of size bounded by $C\delta^{-s/(1-s)}$ with a constant $C > 0$ independent of the dimension of the input data $z$. Similar results for parametric PDEs can be found in [28].

The methods in [36] motivated the work of [19], in which the authors show holomorphy of the data-to-QoI map $y \mapsto \mathbb{E}^{\mu^*}[\mathrm{QoI}]$, which relates observation data to the posterior expectation of an unknown quantity of interest (QoI) for additive centered Gaussian observation noise in Bayesian inverse problems. Using the fact that holomorphy implies fast convergence of Taylor expansions, the authors derived an exponential expression rate bound in terms of the overall network size.

Our approach differs from the ideas above as we do not want to approximate the data-to-QoI map, but instead emulate the state $p$ itself by a DNN. Hence in our method the input of the neural network is a point in the domain of the state, $x \in D$. The output of the neural network is an approximation of the state at this point, $p_\theta(x) \in \mathbb{R}$, i. e., $N_L = 1$. By a slight abuse of notation we denote by $p_\theta \in \mathcal{V} = \mathbb{R}^{n_p}$ also a vector containing evaluations of the neural network at the $n_p$-many grid points of the state. In combination with a one-shot approach for the training of the neural network parameters, our method is closer related to the physics-informed neural networks (PINNs) in [30, 31].

In [30, 31] the authors consider PDEs of the form

$$f(t,x) := p_t + N(p,\lambda) = 0, \quad t \in [0,T],\ x \in D,$$

where $N$ is a nonlinear differential operator parameterized by $\lambda$. The authors replace $p$ by a neural network $p_\theta$ and use automatic differentiation to construct the function $f_\theta(t,x)$. The neural network parameters are then obtained by minimizing $\mathrm{MSE} = \mathrm{MSE}_p + \mathrm{MSE}_f$, where

$$\mathrm{MSE}_p := \frac{1}{N_p}\sum_{i=1}^{N_p}\left|p_\theta(t_p^i, x_p^i) - p^i\right|^2, \quad \mathrm{MSE}_f := \frac{1}{N_f}\sum_{i=1}^{N_f}\left|f_\theta(t_f^i, x_f^i)\right|^2,$$

$\{t_p^i, x_p^i, p^i\}_{i=1}^{N_p}$ denote the training data, and $\{t_f^i, x_f^i\}_{i=1}^{N_f}$ are collocation points of $f_\theta(t,x)$. For the minimization, an L-BFGS method is used. The parameters $\lambda$ of the differential operator turn into parameters of the neural network $f_\theta$ and can be learned by minimizing the MSE.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

In [39] the authors consider the so-called Bayesian neural networks (BNNs), where the neural network parameters are updated according to Bayes' theorem. Hereby the initial distribution on the network parameters serves as prior distribution. The likelihood requires the PDE solution, which is obtained by concatenating the Bayesian neural network with a physics-informed neural network, which they call Bayesian physics-informed neural networks (B-PINNs). For the estimation of the posterior distributions, they use the Hamiltonian Monte Carlo method and variational inference. In contrast to the PINNs, the Bayesian framework allows them to quantify the aleatoric uncertainty associated with noisy data. In addition to that, their numerical experiments indicate that B-PINNs beat PINNs in case of large noise levels on the observations. In contrast to that, our proposed method is based on the MAP estimate and remains exact in the small noise limit. We propose a derivative-free optimization method, the EKI, which shows promising results (also compared to quasi-Newton methods) without requiring derivatives with respect to the weights and design parameters.

## 14.4 Ensemble Kalman filter for neural network-based one-shot inversion

The ensemble Kalman inversion (EKI) generalizes the well-known ensemble Kalman filter (EnKF) introduced by Evensen and coworker in the data assimilation context [14] to the inverse setting; see [21] for more detail. Since the Kalman filter involves a Gaussian approximation of the underlying posterior distribution, we focus on an iterative version based on tempering to reduce the linearization error. Recall the posterior distribution $\mu^*$ given by

$$\mu^*(\mathrm{d}v) \propto \exp\left(-\frac{1}{2}\|G(v) - y\|_\Gamma^2\right)\mu_0(\mathrm{d}v)$$

for an abstract inverse problem

$$y = G(v) + \eta$$

with $G$ mapping from the unknowns $v \in \mathbb{R}^{n_v}$ to the observations $y \in \mathbb{R}^{n_y}$ with $\eta \sim \mathcal{N}(0, \Gamma)$, $\Gamma \in \mathbb{R}^{n_y \times n_y}$ s. p. d. We define the intermediate measures

$$\mu_n(\mathrm{d}v) \propto \exp\left(-\frac{1}{2}nh\|G(v) - y\|_\Gamma^2\right)\mu_0(\mathrm{d}v), \quad n = 0, \dots, N, \tag{14.9}$$

by scaling the data misfit/likelihood by the step size $h = N^{-1}$, $N \in \mathbb{N}$. The idea is to evolve the prior distribution $\mu_0$ into the posterior distribution $\mu_N = \mu^*$ by this sequence of intermediate measures and to apply the EnKF to the resulting artificial time

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

dynamical system. Note that we account for the repeated use of the observations by amplifying the noise variance by $N = 1/h$ in each step. Then the EKI uses an ensemble of $J$ particles $\{v_0^{(j)}\}_{j=1}^J$ with $J \in \mathbb{N}$ to approximate the intermediate measures $\mu_n$ by

$$\mu_n \simeq \frac{1}{J}\sum_{j=1}^{J} \delta_{v_n}^{(j)}$$

with $\delta_v$ denoting the Dirac measure centered at $v_n^{(j)}$. The particles are transformed in each iteration by the application of the Kalman update formulas to the empirical mean $\bar{v}_n = \frac{1}{J}\sum_{j=1}^J v_n^{(j)}$ and covariance $C(v_n) = \frac{1}{J-1}\sum_{j=1}^J (v_n^{(j)} - \bar{v}_n) \otimes (v_n^{(j)} - \bar{v}_n)$ in the form

$$\bar{v}_{n+1} = \bar{v}_n + K_n\big(y - G(\bar{v}_n)\big), \qquad C(v_{n+1}) = C(v_n) - K_n C^{y,v}(v_n),$$

where $K_n = C^{v,y}(v_n)(C^{y,y}(v_n) + \frac{1}{h}\Gamma)^{-1}$ denotes the Kalman gain, and for $v = \{v^{(j)}\}_{j=1}^J$, the operators $C^{yy}$ and $C^{vy}$ given by

$$\begin{aligned}
C^{y,y}(v) &= \frac{1}{J}\sum_{j=1}^{J} \big(G(v^{(j)}) - \bar{G}\big) \otimes \big(G(v^{(j)}) - \bar{G}\big),\\
C^{v,y}(v) &= \frac{1}{J}\sum_{j=1}^{J} \big(v^{(j)} - \bar{v}\big) \otimes \big(G(v^{(j)}) - \bar{G}\big),\\
C^{y,v}(v) &= \frac{1}{J}\sum_{j=1}^{J} \big(G(v^{(j)}) - \bar{G}\big) \otimes \big(v^{(j)} - \bar{v}\big),\\
\bar{G} &= \frac{1}{J}\sum_{j=1}^{J} G(v^{(j)})
\end{aligned}$$

are the empirical covariances and means in the observation space. Since this update does not uniquely define the transformation of each particle $v_n^{(j)}$ to the next iteration $v_{n+1}^{(j)}$, the specific choice of the transformation leads to different variants of the EKI. We focus here on the generalization of the EnKF as introduced by [21], resulting in a mapping of the particles of the form

$$v_{n+1}^{(j)} = v_n^{(j)} + C^{v,y}(v_n)\big(C^{y,y}(v_n) + h^{-1}\Gamma\big)^{-1}\big(y_{n+1}^{(j)} - G(v_n^{(j)})\big), \quad j = 1, \dots, J, \tag{14.10}$$

where

$$y_{n+1}^{(j)} = y + \xi_{n+1}^{(j)}.$$

The $\xi_{n+1}^{(j)}$ are i. i. d. random variables distributed according to $\mathcal{N}(0, h^{-1}\Sigma)$ with $\Sigma = \Gamma$ corresponding to the case of perturbed observations and $\Sigma = 0$ to the unperturbed observations.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

The motivation via the sequence of intermediate measures and the resulting artificial time allows us to derive the continuous time limit of the iteration, which has been extensively studied in [4, 33, 34] to build analysis of the EKI in the linear setting. This limit arises by taking the parameter $h$ in (14.10) to zero, resulting in

$$\frac{\mathrm{d}v^{(j)}}{\mathrm{d}t} = C^{v,y}(v)\Gamma^{-1}\big(y - G(v^{(j)})\big) + C^{v,y}(v^{(j)})\Gamma^{-1}\sqrt{\Sigma}\,\frac{dW^{(j)}}{dt}.$$

As shown in [13], the EKI does not in general converge to the true posterior distribution. Therefore the analysis presented in [4, 33, 34] views the EKI as a derivative-free optimizer of the data misfit, which is also the viewpoint we adopt here.

### 14.4.1 Ensemble Kalman inversion for neural network-based one-shot formulation

By approximating the state of the underlying PDE by a neural network we seek to optimize the unknown parameter $u$ and on the other side the parameters of the neural network $\theta$. The idea is based on defining the function $H(v) := H(u,\theta) = F(u, p_\theta)$, where $p_\theta$ denotes the state approximated by the neural network and $v = (u,\theta)^\top$. This leads to the empirical summary statistics

$$\begin{aligned}
\overline{(u,\theta)}_n &= \frac{1}{J}\sum_{j=1}^{J}(u_n^{(j)}, \theta_n^{(j)}), \quad \bar{H}_n = \frac{1}{J}\sum_{j=1}^{J} H(u_n^{(j)}, \theta_n^{(j)}),\\
C_n^{u\theta,y} &= \frac{1}{J}\sum_{j=1}^{J}\big((u_n^{(j)}, \theta_n^{(j)})^\top - \overline{(u,\theta)}_n^\top\big) \otimes \big(H(u_n^{(j)}, \theta_n^{(j)}) - \bar{H}_n\big),\\
C_n^{y,y} &= \frac{1}{J}\sum_{j=1}^{J}\big(H(u_n^{(j)}, \theta_n^{(j)}) - \bar{H}_n\big) \otimes \big(H(u_n^{(j)}, \theta_n^{(j)}) - \bar{H}_n\big),
\end{aligned}$$

and the EKI update

$$(u_{n+1}^{(j)}, \theta_{n+1}^{(j)})^\top = (u_n^{(j)}, \theta_n^{(j)})^\top + C_n^{u\theta,y}\big(C_n^{y,y} + h^{-1}\Gamma\big)^{-1}\big(\tilde{y}_{n+1}^{(j)} - H(u_n^{(j)}, \theta_n^{(j)})\big), \tag{14.11}$$

where the perturbed observation is computed as before:

$$\tilde{y}_{n+1}^{(j)} = \tilde{y} + \xi_{n+1}^{(j)}, \quad \xi_{n+1}^{(j)} \sim \mathcal{N}(0, h^{-1}\Sigma),$$

with

$$\tilde{y} = \begin{pmatrix} 0 \\ y \end{pmatrix}, \quad \Gamma := \begin{pmatrix} \Gamma_{\text{model}} & 0 \\ 0 & \Gamma_{\text{obs}} \end{pmatrix}.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Figure 14.1 illustrates the basic idea of the application of the EKI to solve the neural network-based one-shot formulation.

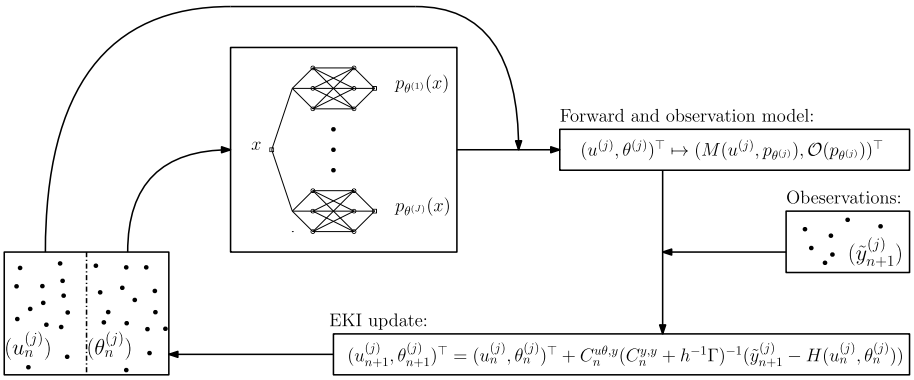



**Figure 14.1:** Description of the EKI applied to solve the neural network-based one-shot formulation.

The EKI (14.11) will be used as a derivative-free optimizer of the data misfit $\|F(u, p_\theta) - \tilde{y}\|_\Gamma^2$. The analysis presented in [4, 33, 34] shows that the EKI in its continuous form is able to recover the data with a finite number of particles in the limit as $t \to \infty$ under suitable assumptions on the forward problem and the set of particles. In particular, the analysis assumes a linear forward problem. Extensions to the nonlinear setting can be found, e. g., in [7]. The limit as $t \to \infty$ corresponds to the noise-free setting, as the inverse noise covariance scales with $n/N = nh$ in (14.9). To explore the scaling of the noise and to discuss regularization techniques, we illustrate the ideas in the following for a Gaussian linear setting, i. e., we assume that the forward response operator is linear $H(v) = Av$ with $A \in \mathcal{L}(\mathcal{X} \times \Theta, \mathbb{R}^{n_w + n_y})$ and $\mu_0 = \mathcal{N}(v_0, C_0)$. Considering the large ensemble size limit as $J \to \infty$, the mean $m$ and covariance $C$ satisfy the equations

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} = -C(t)A^\top \Gamma^{-1}(Am(t) - y),$$
$$\frac{\mathrm{d}C}{\mathrm{d}t} = -C(t)A^\top \Gamma^{-1} AC(t)$$

for $\Sigma = \Gamma$ in (14.10). By considering the dynamics of the inverse covariance it is straightforward to show that the solution is given by

$$C^{-1}(t) = C_0^{-1} + A^\top \Gamma^{-1} At;$$

see, e. g., [15] and the references therein for details. Note that $C(1)$ corresponds to the posterior covariance and that $C(t) \to 0$ as $t \to \infty$. Furthermore, the mean is given by

$$m(t) = \left(C_0^{-1} + A^\top \Gamma^{-1} At\right)^{-1}\left(A^\top \Gamma^{-1} yt + C_0^{-1} v_0\right);$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

in particular, the mean minimizes the data misfit in the limit as $t \to \infty$. Therefore the application of the EKI in the inverse setting often requires additional techniques such as adaptive stopping [34] or additional regularization [6] to overcome the ill-posedness of the minimization problem. To control the regularization of the data misfit and neural network individually, we consider the following system:

$$F(u, p_\theta) + \begin{pmatrix} \eta_{\text{model}} \\ \eta_{\text{obs}} \end{pmatrix} = \tilde{y},$$
$$\begin{pmatrix} u \\ \theta \end{pmatrix} + \begin{pmatrix} \eta_{\text{param}} \\ \eta_{NN} \end{pmatrix} = 0$$

with $\eta_{\text{model}} \sim \mathcal{N}(0, 1/\lambda\, \hat{\Gamma}_{\text{model}})$, $\eta_{\text{obs}} \sim \mathcal{N}(0, \Gamma_{\text{obs}})$, $u \sim \mathcal{N}(u_0, 1/\alpha_1\, C)$, and $\theta \sim \mathcal{N}(0, 1/\alpha_2\, I)$. Therefore the loss function for the augmented system is given by

$$\frac{1}{2}\left\|O(p_\theta) - y\right\|^2_{\Gamma_{\text{obs}}} + \frac{\lambda}{2}\left\|M(u, p_\theta)\right\|^2_{\hat{\Gamma}_{\text{model}}} + \frac{\alpha_1}{2}\|u - u_0\|^2_C + \frac{\alpha_2}{2}\|\theta\|^2. \tag{14.12}$$

Assuming that the resulting forward operator

$$G(u, \theta) = \begin{pmatrix} F(u, p_\theta) \\ u \\ \theta \end{pmatrix} \tag{14.13}$$

is linear, the EKI will converge to the minimum of the regularized loss function (14.12), cf. [33]. To ensure the feasibility of the EKI estimate (withe respect to the underlying forward problem), we propose the following algorithm using the ideas discussed in Section 14.2.5.

**Theorem 14.4.1.** *Assume that the forward operator* $G : \mathcal{X} \times \Theta \to \mathbb{R}^{n_G}$ *where* $n_G := n_w + n_y + n_u + n_\theta$ *and*

$$G(u, \theta) = \begin{pmatrix} F(u, p_\theta) \\ u \\ \theta \end{pmatrix}$$

*is linear, i. e.,* $F(u, p_\theta) = A\binom{u}{\theta}$ *with* $A \in \mathcal{L}(\mathcal{X} \times \Theta, \mathbb{R}^{n_w + n_y})$. *Let* $(\lambda_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$ *be strictly monotonically increasing and* $\lambda_k \to \infty$ *as* $k \to \infty$. *Further, assume that the initial ensemble members are chosen so that* $\operatorname{span}\{(u^{(j)}(0), \theta^{(j)}(0))^\top, j = 1, \dots, J\} = \mathcal{X} \times \Theta$.

*Then Algorithm* 14.1 *generates a sequence of estimates* $(\bar{u}_k, \bar{\theta}_k)_{k\in\mathbb{N}}$, *where* $\bar{u}_k, \bar{\theta}_k$ *minimizes the loss function for the augmented system given by*

$$\frac{1}{2}\left\|O(p_\theta) - y\right\|^2_{\Gamma_{\text{obs}}} + \frac{\lambda_k}{2}\left\|M(u, p_\theta)\right\|^2_{\hat{\Gamma}_{\text{model}}} + \frac{\alpha_1}{2}\|u - u_0\|^2_C + \frac{\alpha_2}{2}\|\theta\|^2$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Algorithm 14.1** Penalty ensemble Kalman inversion for neural network-based one-shot inversion.

**Require:** initial ensemble $v_0^{(j)} = (u_0^{(j)}, \theta_0^{(j)})^\top \in \mathcal{X} \times \Theta, j = 1, \dots J$, $\lambda_0$.

1: **for** $k = 0, 1, 2, \dots$ **do**

2: Compute an approximation of the minimizer $(u_k, \theta_k)^\top$ of

$$\min_{u,\theta} \frac{1}{2}\|O(p_\theta) - y\|^2_{\Gamma_{\text{obs}}} + \frac{\lambda_k}{2}\|M(u, p_\theta)\|^2_{\hat{\Gamma}_{model}} + \frac{\alpha_1}{2}\|u - u_0\|^2_C + \frac{\alpha_2}{2}\|\theta\|^2$$

by solving

$$\frac{\mathrm{d}v^{(j)}}{\mathrm{d}t} = C^{vy}(v)\Gamma^{-1}(\hat{y} - G(v^{(j)})) + C^{vy}(v^{(j)})\Gamma^{-1}\sqrt{\Sigma}\frac{dW^{(j)}}{dt}$$

with $\hat{y} = (0, y, 0, 0)^\top$, $v^{(j)}(0) = v_0^{(j)}$ for system (14.13) and $\Gamma = \text{diag}(C, I, \frac{1}{\alpha_1}C, \frac{1}{\alpha_2}I)$.

3: Set $v_k = (u_k, \theta_k)^\top = \lim_{T\to\infty} \bar{v}(T)$.

4: Increase $\lambda_k$.

5: Draw $J$ ensemble members $v_0^{(j)}$ from $\mathcal{N}(v_k, \left(\begin{smallmatrix} C & 0 \\ 0 & I \end{smallmatrix}\right))$.

6: **end for**

*with given $\alpha_1, \alpha_2 > 0$. Furthermore, every accumulation point of $(\bar{u}_k, \bar{\theta}_k)_{k\in\mathbb{N}}$ is the (unique, global) minimizer of*

$$\min_{u,\theta} \frac{1}{2}\|O(p_\theta) - y\|^2_{\Gamma_{\text{obs}}} + \frac{\alpha_1}{2}\|u - u_0\|^2_C + \frac{\alpha_2}{2}\|\theta\|^2$$
$$s.t.\ M(u, p_\theta) = 0.$$

*Proof.* Under the assumption of a linear forward model, the penalty function

$$\frac{1}{2}\|O(p_\theta) - y\|^2_{\Gamma_{\text{obs}}} + \frac{\lambda_k}{2}\|M(u, p_\theta)\|^2_{\hat{\Gamma}_{\text{model}}} + \frac{\alpha_1}{2}\|u - u_0\|^2_C + \frac{\alpha_2}{2}\|\theta\|^2$$

is strictly convex for all $k \in \mathbb{N}$, i. e., there exists a unique minimizer of the penalized problem. Choosing the initial ensemble such that $\text{span}\{(u^{(j)}(0), \theta^{(j)}(0))^\top, j = 1, \dots, J\} = \mathcal{X} \times \Theta$ ensures the convergence of the EKI estimate to the global minimizer; see [6, Theorem 3.13] and [33, Theorem 4]. The convergence of Algorithm 14.1 to the minimizer of the constrained problem then follows from Proposition 14.2.1. □

**Remark 14.4.1.** Note that the convergence result Theorem 14.4.1 involves an assumption on the size of the ensemble to ensure the convergence to the (global) minimizer of the loss function in each iteration. This is due to the well-known subspace property of the EKI, i. e., the EKI estimate will lie in the span of the initial ensemble when using the EKI in its variant as discussed here. In case of a large or possibly infinite-dimensional

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

parameter/state space, the assumption on the size of the ensemble is usually not satisfied in practice. Techniques such as variance inflation, localization, and adaptive ensemble choice are able to overcome the subspace property and thus might lead to much more efficient algorithms from a computational point of view. Furthermore, we stress the fact that the convergence result presented above requires the linearity of the forward and observation operator (with respect to the optimization variables), i. e., the assumption is not fulfilled when considering neural networks with a nonlinear activation function as approximation of the forward problem. Note that the proposed approach can be applied in a rather general framework for arbitrary approximation methods of the forward problem, i. e., in particular for finite element approximations. However, in the case of neural networks with nonlinear activation functions, the assumptions are not met, but numerical experiments show promising results even in the nonlinear setting. Therefore, Theorem 14.4.1 is included to cover the case with linear approximations of the forward problem and as a starting point for the analysis of nonlinear problems, which will be subject to future work.

To accelerate the computation of the minimizer, we suggest the following variant. Algorithm 14.1 requires the solution of a sequence of optimization problems, i. e., for each $\lambda$, EKI is used to approximate the solution of the corresponding minimization problem. To avoid the repeated application of EKI, the idea of Algorithm 14.2 is to solve just one optimization problem with increasing regularization parameter $\lambda$ (instead of the sequence of optimization problems). This is incorporated in the continuous ver-

**Algorithm 14.2** Simultaneous penalty ensemble Kalman inversion for neural network-based one-shot inversion.

**Require:** initial ensemble $v_0^{(j)} = (u_0^{(j)}, \theta_0^{(j)})^\top \in \mathcal{X} \times \Theta, j = 1, \dots, J$, $\lambda_0 \in \mathbb{R}_{\geq 0}$, $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_+$.

1: Compute an approximation of the minimizer of

$$\min_{u,\theta} \frac{1}{2}\|O(p_\theta) - y\|^2_{\Gamma_{\text{obs}}} + \frac{\alpha_1}{2}\|u - u_0\|^2_C + \frac{\alpha_2}{2}\|\theta\|^2$$
$$\text{s. t. } M(u, p_\theta) = 0$$

by solving the system

$$\frac{\mathrm{d}v^{(j)}}{\mathrm{d}t} = C^{vy}(v)\Gamma^{-1}(\hat{y} - G(v^{(j)})) + C^{vy}(v^{(j)})\Gamma^{-1}\sqrt{\Sigma}\frac{dW^{(j)}}{dt},$$
$$\frac{\mathrm{d}\lambda}{\mathrm{d}t} = f(\lambda),$$

with $\hat{y} = (0, y, 0, 0)^\top$, $v^{(j)}(0) = v_0^{(j)}$ for system (14.13), $\lambda(0) = \lambda_0$, and $\Gamma = \operatorname{diag}(C, I, \frac{1}{\alpha_1}C, \frac{1}{\alpha_2}I)$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

sion of EKI by solving an additional differential equation for $\lambda$ with nondecreasing right-hand side. The computational effort is thus reduced, and numerical experiments suggest a comparable performance in terms of accuracy. The theoretical analysis of the convergence behavior will be subject to future work.

## 14.5 Numerical results

We present in the following numerical experiments illustrating the one-shot inversion. The first example is a one-dimensional problem, for which we compare the black-box method, quasi-Newton method for the one-shot inversion, quasi-Newton method for the neural network-based one-shot inversion (Algorithm 14.1), EKI for the one-shot inversion, and EKI for the neural network-based one-shot inversion (Algorithm 14.2) in the linear setting. Furthermore, we numerically explore the convergence behavior of the EKI for the neural network-based one-shot inversion (Algorithm 14.2) also for a nonlinear forward model. The second experiment is concerned with the extension of the linear model to the two-dimensional problem to investigate the potential of the EKI for neural network-based inversion in the higher-dimensional setting.

### 14.5.1 One-dimensional example

We consider the problem of recovering the unknown data $u^\dagger$ from noisy observations

$$y = \mathcal{O}(p^\dagger) + \eta^\dagger,$$

where $p^\dagger = \mathcal{A}^{-1}(u^\dagger)$ is the solution of the one-dimensional elliptic equation

$$\begin{aligned} -\frac{\mathrm{d}^2 p}{\mathrm{d}x^2} + p &= u^\dagger \quad \text{in } D := (0,\pi), \\ p &= 0 \quad \text{on } \partial D, \end{aligned} \tag{14.14}$$

with operator $\mathcal{O}$ observing the dynamical system at $n_y = 2^3 - 1$ equispaced observation points $x_i = \frac{i}{2^4} \cdot \pi$, $i = 1, \dots, n_y$.

We approximate the forward problem (14.14) numerically on a uniform mesh with meshwidth $h = 2^{-6}$ by a finite element method with continuous piecewise linear ansatz functions. The approximated solution operator will be denoted by $S \in \mathbb{R}^{n_u \times n_u}$ with $n_u = 1/h$.

The unknown parameter $u$ is assumed to be Gaussian, i. e., $u \sim \mathcal{N}(0, C_0)$, with (discretized) covariance operator $C_0 = \beta(-\frac{\mathrm{d}^2}{\mathrm{d}x^2})^{-\nu}$ for $\beta = 5$ and $\nu = 1.5$. For our inverse problem, we assume a observational noise covariance $\Gamma_{\text{obs}} = 0.1 \cdot I_{n_y}$ and a model error covariance $\hat{\Gamma}_{\text{model}} = 100 \cdot I_{n_u}$, and we choose the regularization parameter $\alpha_1 = 0.002$, whereas we turn off the regularization on $p$, i. e., we set $\alpha_2 = 0$. Further, we choose

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

a feed-forward DNN with $L = 3$ layers, where we set $N_1 = N_2 = 10$ the size of the hidden layers and $N_0 = N_L = 1$ the size of the input and output layers. As activation function, we choose the sigmoid function $\varrho(x) = \frac{1}{1+e^{-x}}$. The EKI method is based on the deterministic formulation represented through the coupled ODE system

$$\frac{\mathrm{d}v^{(j)}}{\mathrm{d}t} = C^{vy}(v)\Gamma^{-1}(y - G(v^{(j)})), \tag{14.15}$$

which will be solved with the MATLAB function ode45 up to time $T = 10^{10}$. The ensembles of particles $(u^{(j)})$, $(u^{(j)}, p^{(j)})$, and $(u^{(j)}, \theta^{(j)})$ will be initialized by $J = 150$ particles as i. i. d. samples, where the parameters $u_0^{(j)}$ are drawn from the prior distribution $\mathcal{N}(0, C_0)$, the states $p_0^{(j)}$ are drawn from $\mathcal{N}(0, 5\,I_{n_p})$, and the weights of the neural network approximation are drawn from $\mathcal{N}(0, I_{n_\theta})$, which are all independent from each other.

We compare the results to a classical gradient-based method, namely the quasi-Newton method with BFGS updates, as implemented by MATLAB.

We summarize the methods in the following and introduce abbreviations:

1. reduced formulation: explicit solution (redTik).
2. one-shot formulation: we compare the performance of the EKI with Algorithm 14.1 (osEKI_1), the EKI with Algorithm 14.2 (osEKI_2), and the quasi-Newton method with Algorithm 14.1 (osQN_1).
3. neural network-based one-shot formulation: we compare the performance of the EKI with Algorithm 14.2 (nnosEKI_2) and the quasi-Newton method with Algorithm 14.1 (nnosQN_1).

Figure 14.2 shows the increasing sequence of $\lambda$ used for Algorithm 14.1 and the quasi-Newton method and for Algorithm 14.2 (over time).

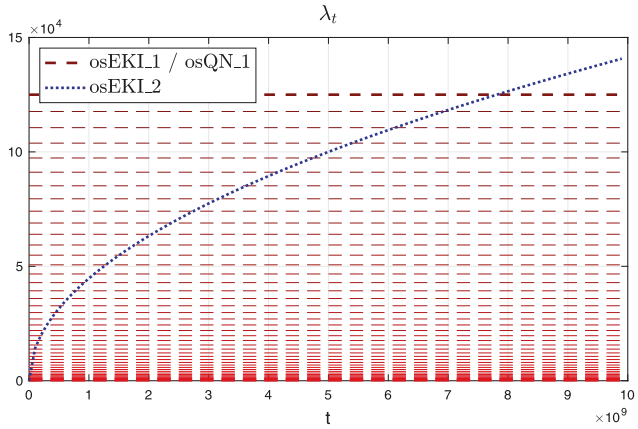



**Figure 14.2:** Scaling parameter $\lambda$ depending on time for Algorithm 14.1, $\lambda_k = k^3$ for $k = 1, 2, \ldots, 50$, and Algorithm 14.2, $\mathrm{d}\lambda/\mathrm{d}t = 1/\lambda$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

#### 14.5.1.1 One-shot inversion

To illustrate the convergence result of the EKI and to numerically investigate the performance of Algorithm 14.2, we start the discussion by a comparison of the one-shot inversion based on the FEM approximation of the forward problem in the 1d example.

Figure 14.3 illustrates the difference of the estimates given by EKI with Algorithm 14.1 (osEKI_1), the EKI with Algorithm 14.2 (osEKI_2) and the quasi-Newton method with Algorithm 14.1 (osQN_1) compared to the Tikhonov solution and the truth (on the left-hand side), and in the observation space (on the right-hand side). We observe that all three methods lead to an excellent approximation of the Tikhonov solution. Due to the linearity of the forward problem, the quasi-Newton method and the EKI with Algorithm 14.1 are expected to converge to the regularized solution. The EKI with Algorithm 14.2 shows a similar performance while reducing the computational effort significantly compared to Algorithm 14.1.

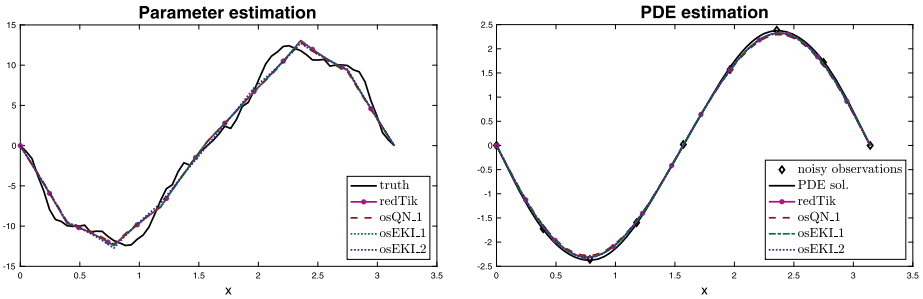



**Figure 14.3:** Comparison of parameter estimation given by EKI with Algorithm 14.1 (osEKI_1), the EKI with Algorithm 14.2 (osEKI_2), and the quasi-Newton method with Algorithm 14.1 (osQN_1) compared to the Tikhonov solution and the truth (on the left-hand side), and in the observation space (on the right-hand side).

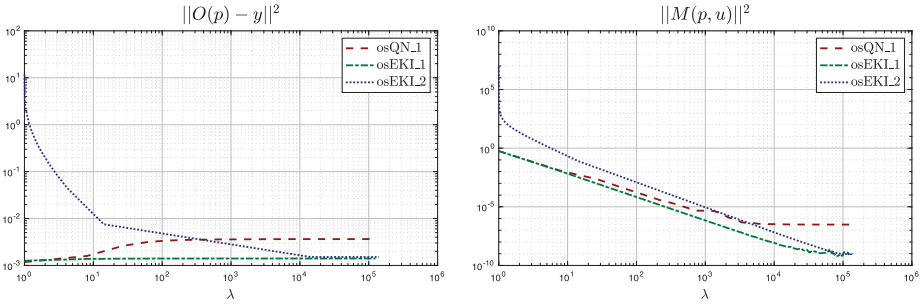



**Figure 14.4:** Comparison of the data misfit given by EKI with Algorithm 14.1 (osEKI_1), the EKI with Algorithm 14.2 (osEKI_2), and the quasi-Newton method with Algorithm 14.1 (osQN_1) (on the left-hand side) and residual of the forward problem (on the right-hand side), both with respect to $\lambda$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

The comparison of the data misfit and the residual of the forward problem shown in Figure 14.4 reveals a very good performance of the EKI (for both algorithms) with feasibility of the estimate (with respect to the forward problem) in the range of $10^{-10}$.

#### 14.5.1.2 One-shot method with neural network approximation

The next experiment replaces the forward problem by a neural network in the one-shot setting. Due to the excellent performance of Algorithm 14.2 in the previous experiment, we focus in the following on this approach for the neural network-based one-shot inversion.

The EKI for the neural network-based one-shot inversion leads to a very good approximation of the regularized solution (see Figure 14.5), whereas the performance of

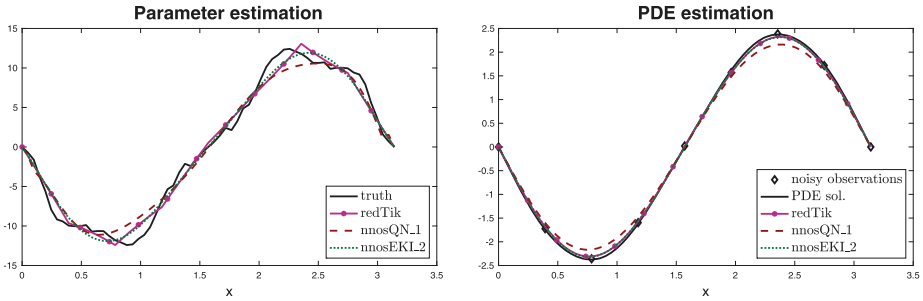



**Figure 14.5:** Comparison of parameter estimation given by the EKI with Algorithm 14.2 (nnosEKI_2) and the quasi-Newton method with Algorithm 14.1 (nnosQN_1) for the neural network-based one-shot inversion compared to the Tikhonov solution and the truth (on the left-hand side) and in the observation space (on the right-hand side).

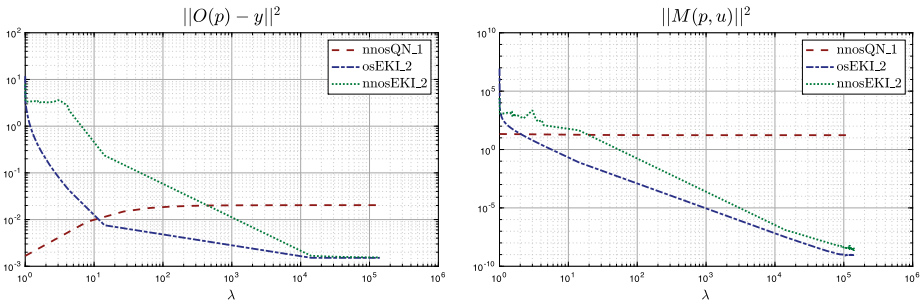



**Figure 14.6:** Comparison of the data misfit given by the EKI with Algorithm 14.2 (nnosEKI_2) and the quasi-Newton method with Algorithm 14.1 (nnosQN_1) for the neural network-based one-shot inversion compared to EKI with Algorithm 14.2 (osEKI_2) from the previous experiment (on the left-hand side) and residual of the forward problem (on the right-hand side), both with respect to $\lambda$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

the quasi-Newton approach is slightly worse, which might be attributed to the nonlinearity introduced by the neural network approximation.

The comparison of the data misfit and residual of the forward problem reveals an excellent convergence behavior of the EKI for the neural network-based one-shot optimization, whereas the quasi-Newton method does not converge to a feasible estimate; see Figure 14.6.

#### 14.5.1.3 Nonlinear forward model

We consider in the following nonlinear forward model of the form

$$\begin{aligned} -\nabla \cdot (\exp(u^\dagger) \cdot \nabla p) &= 10 \quad \text{in } D := (0, \pi), \\ p &= 0 \quad \text{on } \partial D. \end{aligned}$$

Note that the mapping from the unknown parameter function to the state is nonlinear. We use the same discretization as in the linear problem. The unknown parameter $u^\dagger$ is assumed to be Gaussian with zero mean and $C_0 = \beta(-\frac{\mathrm{d}^2}{\mathrm{dx}^2})^{-\nu}$, where we choose $\beta = 1$ and $\nu = 2$. Further, we set $\Gamma_{\text{obs}} = 0.0001 \cdot I_{n_y}$, $\hat{\Gamma}_{\text{model}} = 10 \cdot I_{n_u}$, $\alpha_1 = 2$, and $\alpha_2 = 0$. Furthermore, the structure of the feed-forward DNN remains the same as in the linear case.

We compare the one-shot method with neural network approximation resulting from the EKI with Algorithm 14.2 with the Tikhonov solution of the reduced formulation, which has been approximated by a quasi-Newton method. We determine the scaling parameter $\lambda$ in Algorithm 14.2 by the ODE $\mathrm{d}\lambda/\mathrm{d}t = 1$, i. e. the scaling parameter grows linearly. Similarly to the linear case, we find that the one-shot method with neural network approximation leads to a good approximation of the Tikhonov solution for the reduced model; see Figure 14.7.

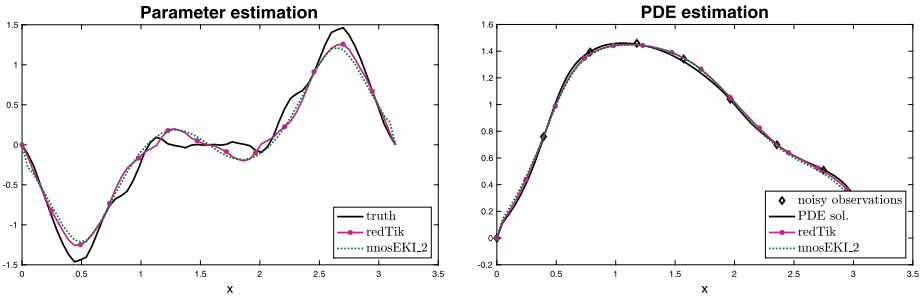



**Figure 14.7:** Comparison of parameter estimation given by the EKI with Algorithm 14.2 (osEKI_2) and the Tikhonov solution (on the left-hand side) and corresponding PDE solution (on the right-hand side).

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

In Figure 14.8, we observe that the penalty parameter $\lambda$ drives the estimate toward feasibility, i. e., toward the solution of the constrained optimization problem.

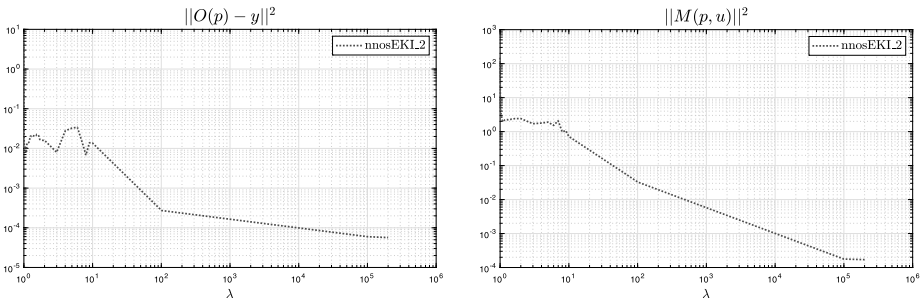



**Figure 14.8:** Data misfit given by the EKI with Algorithm 14.2 (osEKI_2) for the neural network-based one-shot inversion compared (on the left-hand side) and residual of the forward problem (on the right-hand side), both with respect to $\lambda$.

## 14.5.2 Two-dimensional example

Our second numerical example is based on the two-dimensional Poisson equation

$$\begin{aligned} -\Delta p &= u^\dagger \quad \text{in } D := (0,1)^2, \\ p &= 0 \quad \text{on } \partial D, \end{aligned} \tag{14.16}$$

for which we consider again the problem of recovering the unknown source term $u^\dagger$ from noisy observations

$$y = \mathcal{O}(p^\dagger) + \eta^\dagger$$

with $p^\dagger$ denoting the solution of (14.16). We consider an observation operator $\mathcal{O}$ observing $n_y = 50$ randomly picked observation points $x_i$, $i = 1, \ldots, n_y$, shown in Figure 14.9.

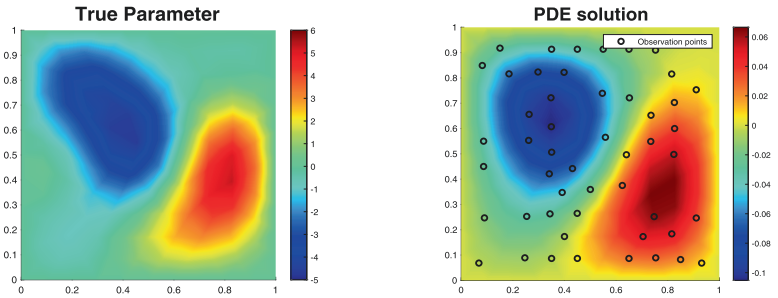



**Figure 14.9:** Ground truth (left-hand side) and the corresponding PDE solution (right-hand side).

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

We numerically approximate the forward model (14.16) with continuous piecewise linear finite element basis functions on a mesh with 95 grid points in $D$ and 40 grid points on $\partial D$ using the `MATLAB Partial Differential Equation Toolbox`. We again denote the approximated solution operator by $S \in \mathbb{R}^{n_u \times n_u}$ with $n_u = 95$. As before, we assume the unknown parameter $u$ to be Gaussian, this time with (discretized) covariance operator $C_0 = \beta(\tau \cdot \mathrm{id} - \Delta)^{-\nu}$ for $\beta = 100$, $\nu = 2$, and $\tau = 1$. The observational noise covariance is assumed to be $\Gamma_{\mathrm{obs}} = 0.01 \cdot I_{n_y}$, whereas we assume the model covariance to be $\hat{\Gamma}_{\mathrm{model}} = 0.1 \cdot I_{n_u}$. We set the regularization parameter $\alpha_1 = 0.002$ and again $\alpha_2 = 0$. The feed-forward DNN consists of $L = 3$ layers with $N_1 = N_2 = 10$ hidden neurons, $N_0 = 2$ input neurons, and $N_L = 1$ output neuron and sigmoid activation function. The setting of the EKI is as described above with $J = 300$ particles drawn as i. i. d. sample from the prior.

Figure 14.9 shows the truth and the corresponding PDE solution.

In the following, we compare the neural network-based one-shot formulation, solved by the EKI with Algorithm 14.2, to the explicit Tikhonov solution of the reduced formulation. The scaling parameter $\lambda$ in Algorithm 14.2 is determined by the ODE $\mathrm{d}\lambda/\mathrm{d}t = 1/\lambda^2$. Figure 14.10 demonstrates that the EKI leads to a comparable solution.

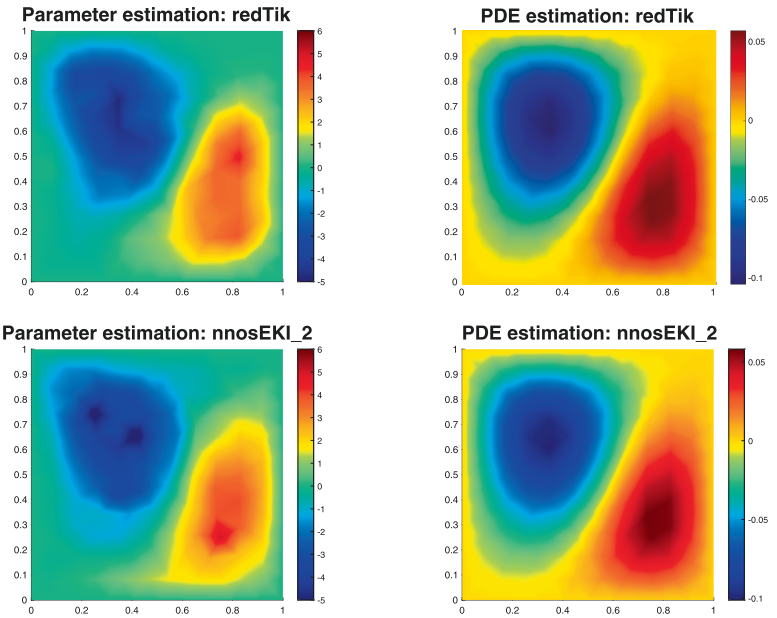



**Figure 14.10:** Comparison of parameter estimation given by the EKI with Algorithm 14.2 (osEKI_2) (below) and the Tikhonov solution (above) (on the left-hand side) and corresponding PDE solution (on the right-hand side).

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

The proposed approach leads to a feasible solution with respect to the forward problem; see Figure 14.11.

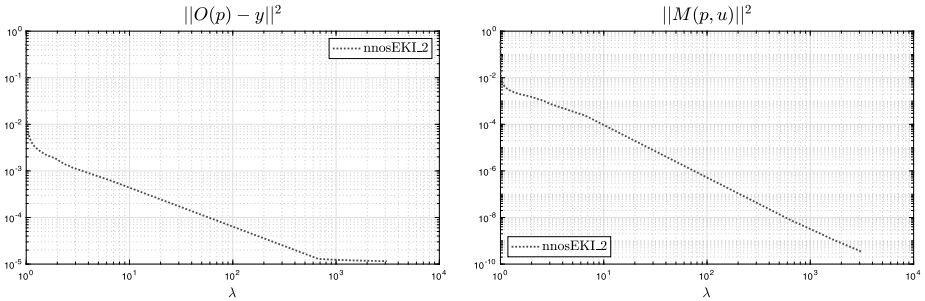



**Figure 14.11:** Data misfit given by the EKI with Algorithm 14.2 (osEKI_2) for the neural network-based one-shot inversion compared (on the left-hand side) and residual of the forward problem (on the right-hand side), both with respect to $\lambda$.

## 14.6 Conclusions

We have demonstrated that the ensemble Kalman inversion for neural network-based one-shot inversion is a promising method, regarding both estimation quality of the unknown parameter and computational feasibility. The link to the Bayesian setting with vanishing noise allowed us to establish a convergence result in a simplified linear setting. Several directions for future work naturally arise from the presented ideas: theoretical analysis of the neural network-based one-shot inversion using recent approximation results of neural networks in the PDE setting, in particular, the comparison of the efficiency of neural network and state-of-the-art approximations of the forward model, parallelization strategies for the EKI using localization ideas, comparison to state-of-the-art optimization methods in the machine learning community.

## Bibliography

[1] S. Agapiou, M. Burger, M. Dashti, and T. Helin. Sparsity-promoting and edge-preserving maximum a posteriori estimators in non-parametric Bayesian inverse problems. *Inverse Probl.*, 34(4):045002, Feb 2018.

[2] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb. Solving inverse problems using data-driven models. *Acta Numer.*, 28:1–174, 2019.

[3] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

[4] D. Blömker, C. Schillings, P. Wacker, and S. Weissmann. Well posedness and convergence analysis of the ensemble Kalman inversion. *Inverse Probl.*, 35(8):085007, Jul 2019.

[5] A. Borzi and V. Schulz. *Computational Optimization of Systems Governed by Partial Differential Equations*. Society for Industrial and Applied Mathematics, USA, 2012.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[6] N. K. Chada, A. M. Stuart, and X. T. Tong. Tikhonov regularization within ensemble Kalman inversion. *SIAM J. Numer. Anal.*, 58(2):1263–1294, 2019.
[7] N. K. Chada and X. T. Tong. Convergence acceleration of ensemble Kalman inversion in nonlinear settings. *arXiv preprint* arXiv:1911.02424, 2019.
[8] C. Clason, T. Helin, R. Kretschmann, and P. Piiroinen. Generalized modes in Bayesian inverse problems. *SIAM/ASA J. Uncertain. Quantificat.*, 7(2):652–684, 2019.
[9] M. Dashti, K. J. H. Law, A. M. Stuart, and J. Voss. MAP estimators and their consistency in Bayesian nonparametric inverse problems. *Inverse Probl.*, 29(9):095017, Sep 2013.
[10] M. Dashti and A. M. Stuart. *The Bayesian Approach to Inverse Problems*, pages 311–428. Springer International Publishing, Cham, 2017.
[11] D. M. Elbrächter, J. Berner, and P. Grohs. *How Degenerate Is the Parametrization of Neural Networks with the ReLU Activation Function?*, pages 7790–7801. Curran Associates, Inc., 2019.
[12] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Mathematics and Its Applications. Springer Netherlands, 1996.
[13] O. G. Ernst, B. Sprungk, and H.-J. Starkloff. Analysis of the ensemble and polynomial chaos Kalman filters in Bayesian inverse problems. *SIAM/ASA J. Uncertain. Quantificat.*, 3(1):823–851, 2015.
[14] G. Evensen. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dyn.*, 53(4):343–367, Nov 2003.
[15] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart. Interacting Langevin diffusions: gradient structure and ensemble Kalman sampler. *SIAM J. Appl. Dyn. Syst.*, 19(1):412–441, 2020.
[16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.
[17] E. Haber, F. Lucka, and L. Ruthotto. Never look back – a modified EnKF method and its application to the training of neural networks without back propagation. *arXiv preprint* arXiv:1805.08034, 2018.
[18] T. Helin and M. Burger. Maximum a posteriori probability estimates in infinite-dimensional Bayesian inverse problems. *Inverse Probl.*, 31(8):085009, Jul 2015.
[19] L. Herrmann, C. Schwab, and J. Zech. Deep ReLU neural network expression rates for data-to-QoI maps in Bayesian PDE inversion. Technical Report 2020-02, Seminar for Applied Mathematics, ETH Zürich, Switzerland, 2020.
[20] D. Higdon, M. Kennedy, J. C. Cavendish, J. A. Cafeo, and R. D. Ryne. Combining field data and computer simulations for calibration and prediction. *SIAM J. Sci. Comput.*, 26(2):448–466, Feb 2005.
[21] M. A. Iglesias, K. J. H. Law, and A. M. Stuart. Ensemble Kalman methods for inverse problems. *Inverse Probl.*, 29(4):045001, 2013.
[22] J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*, volume 160 of Applied Mathematical Sciences. Springer Science & Business Media, New York, NY, 2010.
[23] B. Kaltenbacher. Regularization based on all-at-once formulations for inverse problems. *SIAM J. Numer. Anal.*, 54:2594–2618, 2016.
[24] B. Kaltenbacher. All-at-once versus reduced iterative methods for time dependent inverse problems. *Inverse Probl.*, 33(6):064002, May 2017.
[25] B. Kaltenbacher, A. Neubauer, and O. Scherzer. *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*. Radon Series on Computational and Applied Mathematics, 2008.
[26] M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *J. R. Stat. Soc., Ser. B, Stat. Methodol.*, 63(3):425–464, 2001.
[27] N. B. Kovachki and A. M. Stuart. Ensemble Kalman inversion: a derivative-free technique for machine learning tasks. *Inverse Probl.*, 35(9):095005, Aug 2019.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[28] G. Kutyniok, P. Petersen, M. Raslan, and R. Schneider. A theoretical analysis of deep neural networks and parametric PDEs. *arXiv preprint* arXiv:1904.00377, 2019.
[29] J. A. A. Opschoor, P. C. Petersen, and C. Schwab. Deep ReLU networks and high-order finite element methods. *Anal. Appl.*, 18(5):715–770, 2020.
[30] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics informed deep learning (part I): data-driven solutions of nonlinear partial differential equations. *arXiv preprint* arXiv:1711.10561, 2017.
[31] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics informed deep learning (part II): data-driven discovery of nonlinear partial differential equations. *arXiv preprint* arXiv:1711.10566, 2017.
[32] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.*, 378:686–707, Feb 2019.
[33] C. Schillings and A. M. Stuart. Analysis of the ensemble Kalman filter for inverse problems. *SIAM J. Numer. Anal.*, 55(3):1264–1290, 2017.
[34] C. Schillings and A. M. Stuart. Convergence analysis of ensemble Kalman inversion: the linear, noisy case. *Appl. Anal.*, 97(1):107–123, 2018.
[35] C. Schwab. QMC Galerkin discretization of parametric operator equations. In J. Dick, F. Y. Kuo, G. W. Peters, and I. H. Sloan, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2012*, pages 613–629. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.
[36] C. Schwab and J. Zech. Deep learning in high dimension: neural network expression rates for generalized polynomial chaos expansions in UQ. *Anal. Appl.*, 17(01):19–55, 2019.
[37] Y. Shin, J. Darbon, and G. E. Karniadakis. On the convergence and generalization of physics informed neural networks. *arXiv preprint* arXiv:2004.01806, 2020.
[38] A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numer.*, 19:451–559, 2010.
[39] L. Yang, X. Meng, and G. E. Karniadakis. B-PINNS: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data. *J. Comput. Phys.*, 425:109913, 2021.
[40] D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Netw.*, 94:103–114, 2017.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Joost A. A. Opschoor, Christoph Schwab, and Jakob Zech

# 15 Deep learning in high dimension: ReLU neural network expression for Bayesian PDE inversion

**Abstract:** We establish dimension-independent expression rates by deep ReLU networks for certain countably parametric maps, so-called $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphic functions. These are mappings from $[-1,1]^{\mathbb{N}} \to \mathcal{X}$, with $\mathcal{X}$ being a Banach space, that admit analytic extensions to certain polyellipses in each of the input variables. Parametric maps of this type occur in uncertainty quantification for partial differential equations with uncertain inputs from function spaces, upon the introduction of bases. For such maps, we prove (constructive) expression rate bounds by families of deep neural networks, based on multilevel polynomial chaos expansions. We show that $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphy implies summability and sparsity of coefficients in generalized polynomial chaos expansions. This, in turn, implies deep neural network expression rate bounds.

We apply the results to Bayesian inverse problems for partial differential equations with distributed uncertain inputs from Banach spaces. Our results imply the existence of "neural Bayesian posteriors" emulating the posterior densities with expression rate bounds that are free from the curse of dimensionality and limited only by sparsity of certain gPC expansions. We prove that the neural Bayesian posteriors are robust in large data or small noise asymptotics (e. g., [42]), which can be emulated in a noise-robust fashion.

**Keywords:** Bayesian inverse problems, generalized polynomial chaos, deep networks, uncertainty quantification

**MSC 2010:** 62F15, 65N21, 62M45, 68Q32, 41A25

**Acknowledgement:** Supported in part by SNSF Grant No. 159940. Work performed in part during visit of CS and JZ at the CRM, Montreal, Canada, in March 2019, and by a visit of JZ to the FIM, Department of Mathematics, ETH Zürich, in October 2019. JZ acknowledges support by the Swiss National Science Foundation under Early Postdoc Mobility Fellowship 184530. This paper was written during the postdoctoral stay of JZ at MIT.
The extended version [61] of this work additionally contains appendices with several proofs of results from the main text.

**Joost A. A. Opschoor, Christoph Schwab,** ETH Zürich, SAM, ETH Zentrum, HG G57.1, CH8092 Zürich, Switzerland, e-mails: joost.opschoor@sam.math.ethz.ch, christoph.schwab@sam.math.ethz.ch
**Jakob Zech,** University of Heidelberg, 69120 Heidelberg, Germany, e-mail: jakob.zech@uni-heidelberg.de

https://doi.org/10.1515/9783110695984-015

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

## 15.1 Introduction

The efficient numerical approximation of solution (manifolds) to parameter-dependent partial differential equations (PDEs) has seen significant progress in recent years. We refer, for instance, to [13, 14]. Similarly, and closely related, the treatment of Bayesian inverse problems for well-posed partial (integro-)differential equations with uncertain input data has drawn considerable attention; see, e. g., [21] and the references therein. This is, in part, due to the need to efficiently assimilate noisy observation data into predictions subject to constraints given by certain physical laws governing responses of systems of interest. We mention here the surveys [21, 68] and the references therein. In the present paper, we mathematically study the ability of deep neural networks to express Bayesian posterior probability measures subject to given data and PDE constraints. To this end, we work in an abstract setting accommodating PDE constrained Bayesian inverse problems with function space priors as exposed, e. g., in [21, 39] and the references therein.

Several concrete constructions of function space prior probability measures for Bayesian PDE inversion beyond Gaussian measures on separable Hilbert spaces have been advocated in recent years. We mention in particular the so-called *Besov prior measures* [20, 44].

Recently, several proposals have been put forward advocating the use of DNNs for Bayesian PDE inversion from noisy data; we refer to [9, 40, 79]. These references computationally found good numerical efficiency for DNN expression with various architectures of DNNs. Regarding Deep NNs for "learning" solution maps of PDEs, we mention [75, 79]. Expressive power (approximation) rate bounds for solution manifolds of PDEs were obtained in [74]; results in this reference are also a key in the present analysis of DNN expression of Bayesian posteriors. Specifically, we quantify uncertainty in PDE inversion conditional on noisy observation data using the Bayesian framework. Particular attention is on general convex priors on uncertain function space inputs [21, 39].

The Bayesian approach can incorporate most, if not all, uncertainties of engineering interest in PDE inversion and in graph-based data classification in a systematic manner.

Computational UQ for PDEs poses three challenges: large-scale forward problems need to be solved, high-dimensional parameter spaces arise in parameterization of distributed uncertain inputs (from Banach spaces), and numerical approximation needs to scale favorably in the presence of "big data", resulting in consistent posteriors in the sense of Diaconis and Freedman [23].

Foundational mathematical developments on the question of universality of NNs are, e. g., in [7, 8, 27, 37, 38]. In recent years, so-called *deep neural networks* (DNNs) have seen rapid development and successful deployment in a wide range of applications. Evidence for the benefits afforded by depth of NNs on their expressive power has been documented computationally in an increasing number of applications (see,

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

e. g., [40, 47, 48, 63, 67, 79, 84] and the references therein). The results reported in these references are mostly computational and address particular applications. Independent of these numerical experiments exploring the performance of DNN-based algorithms, the *approximation theory of DNNs* has also advanced in recent years. Distinct from earlier, universality results, e. g., in [7, 8, 37, 38], emphasis in more recent mathematical developments has been on approximation (i. e., "expression") rate bounds for specific function classes and particular DNN architectures. We mention only [10, 60, 64] and the references therein. In [74], we proved that ReLU DNNs can express high-dimensional parametric solution families of elliptic PDEs at rates free from the curse of dimensionality.

Specifically, we adopt the infinite-dimensional formulation of Bayesian inverse problems from [76] and its extensions to general convex prior measures on input function spaces as presented in [39]. Assuming an affine representation system on the uncertain input data, we adopt uniform prior measures on the parameters in the representation.

We prove that ReLU DNNs allow for expressing the parameter-to-response map and the Bayesian posterior density at rates which are determined only by the size of the domains of holomorphy.

### 15.1.1 Recent mathematical results on expressive power of DNNs

Fundamental universality results (amounting essentially to statements on density of shallow NN expressions) on DNN expression in the class of continuous functions have been established in the 1990s (see [65] for a proof and review of results); in recent years, *expression rate bounds* for approximation by DNNs for specific classes of functions have been in the focus of interest. We mention in particular [29] and [10]. There it is shown that deep NNs with a particular architecture allow for approximation rate bounds analogous to those of rather general multiresolution systems when measured in terms of the number $N$ of units in the DNN.

In [18], *convolutional DNNs* were shown capable of expressing multivariate functions given in so-called *hierarchical tensor formats*, a numerical representation inspired by electron structure calculations in computational quantum chemistry.

In [49, 80], ReLU DNNs were shown to be able to express general uni- and multivariate polynomials on bounded domains with uniform accuracy $\delta > 0$ and with complexity (i. e., with the numbers of NN layers and NN units and with nonzero weights) scaling polylogarithmically with respect to $\delta$. The results in [49, 80] allow transferring approximation results from high-order finite and spectral element methods, in particular, exponential convergence results, to certain types of DNNs.

In [69], DNN expression rates for multivariate polynomials were investigated without reference to function spaces. Expression rate bounds explicit in the number of variables and the polynomial degree by deep NNs were obtained. The proofs in [69]

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

strongly depend on a large number of bounded derivatives of the activation function and do not cover the presently considered case of ReLU DNNs.

In [74] we proved dimension-independent DNN expression rate bounds on functions of countably many variables. In [74] we used, as we do in part of the present paper, approximation rate bounds for the $N$-term truncated so-called *generalized polynomial chaos expansions* of the parametric function. These have been investigated thoroughly in recent years (e. g., [3, 4, 15, 16] and the references therein). However, for the present analysis, we require more specific information of polynomial degree distributions in $N$-term approximate gPC expansions as the dimension of the space of active parameters increases. This was investigated by some of the authors recently in [82, 83]. In the present paper, we also draw upon results in these references.

In [54] the authors provided an analysis of expressive power of DNNs for a specific class of multiparametric maps, which have a defined (assumed known) *compositional structure*: they are obtained as (repeated) composition of a possibly large number of simpler functions, depending only on a few variables at a time. It was shown that such functions can be expressed with DNNs at complexity bounded by the dimensionality of constituent functions in the composition and the size of the connectivity graph, thereby alleviating the curse of dimensionality for this class.

### 15.1.2 Contributions

We extend our previous work [74] on ReLU NN expression bounds of countably parametric solution families and QoIs for PDEs with affine-parametric uncertain input. In a first main result, Theorem 15.4.9 of Section 15.4.4, we prove bounds on the expressive power of ReLU DNNs for multiparametric response functions from Bayesian inverse UQ for PDEs and more general operator equations subject to infinitely parametric, uncertain, and "invisible" (i. e., not directly observable) input data. As in [74], we assume that the input-to-solution map has holomorphic dependence on possibly an infinite number of parameters. We have in mind in particular (boundary, eigenvalue, control, etc.) problems for elliptic or parabolic PDEs with uncertain coefficients. These may stem from, for example, domains of definition with uncertain geometry (see, e. g., [17, 41, 47, 66]) in diffusion, incompressible flow, or time-harmonic electromagnetic scattering (see, e. g., [41]). Adopting a countable representation system renders uncertain countably parametric inputs and implies likewise countably parametric output families ("solution manifolds", "response surfaces") of the model under consideration.

In [74], expressive power estimates for deep ReLU NNs for countably parametric solution manifolds were obtained among others for linear second-order elliptic PDEs with uncertain coefficients in the divergence form. Theorems 15.4.9 and 15.5.2 extend [74] in two regards. Firstly, we require merely $(\boldsymbol{b}, \varepsilon)$-holomorphy on polyellipses, rather than on polydiscs as assumed in [74]. This requires essential modifications of the DNN

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

expression rate analysis in [74], as Legendre polynomial chaos expansions are used rather than Taylor expansions. Secondly, we generalize our result from [74] to parametric PDEs posed on a polytopal physical domain D of dimension $d \geq 2$ (instead of $d = 1$).

In the *Bayesian setting* (see [21, 39, 76] and the references therein), it has been shown in [24, 72] that $(\boldsymbol{b}, \varepsilon)$-holomorphy of the QoI is inherited by the Bayesian posterior density if it exists. In the present paper, we analyze expression rates of ReLU DNNs for countably parametric Bayesian posterior densities, which arise from PDE inversion subject to noisy data. We show, in particular, extending our analysis [74], that ReLU DNNs afford expression of such densities at dimension-independent rates. The expression rate bounds are, to a large extent, abstracted from particular model PDEs and apply to a wide class of PDEs and inverse problems (e. g., elliptic and parabolic linear PDEs with uncertain coefficients, domains, or source terms). We also provide in Section 15.6.3 novel bounds on the posterior consistency of the DNN emulated Bayesian posterior in the presently considered general setting.

We refer to [79] for a possible computational approach and detailed numerical experiments for a second-order divergence form PDE with log-Gaussian diffusion coefficient.

### 15.1.3 Notation

We adopt standard notation, consistent with our previous works [82, 83]: $\mathbb{N} = \{1, 2, \dots\}$, $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$, and $\mathbb{R}_+ := \{x \in \mathbb{R} : x \geq 0\}$. The symbol $C$ will stand for a generic positive constant independent of any asymptotic quantities in an estimate and may change its value even within the same equation.

In statements about (generalized) polynomial chaos expansions, we require multiindices $\boldsymbol{\nu} = (\nu_j)_{j\in\mathbb{N}} \in \mathbb{N}_0^{\mathbb{N}}$. The *total order* of a multiindex $\boldsymbol{\nu}$ is denoted by $|\boldsymbol{\nu}|_1 := \sum_{j\in\mathbb{N}} \nu_j$. For the countable set of "finitely supported" multiindices, we write

$$\mathcal{F} := \{\boldsymbol{\nu} \in \mathbb{N}_0^{\mathbb{N}} : |\boldsymbol{\nu}|_1 < \infty\}.$$

Here $\operatorname{supp} \boldsymbol{\nu} = \{j \in \mathbb{N} : \nu_j \neq 0\}$ denotes the *support* of the multiindex $\boldsymbol{\nu}$. The size of the support of $\boldsymbol{\nu} \in \mathcal{F}$ is $|\boldsymbol{\nu}|_0 = \#(\operatorname{supp} \boldsymbol{\nu})$; it will, subsequently, indicate the number of active coordinates in the multivariate monomial term $\boldsymbol{y}^{\boldsymbol{\nu}} := \prod_{j\in\mathbb{N}} y_j^{\nu_j}$.

A subset $\Lambda \subseteq \mathcal{F}$ is called *downward closed*[1] if $\boldsymbol{\nu} = (\nu_j)_{j\in\mathbb{N}} \in \Lambda$ implies $\boldsymbol{\mu} = (\mu_j)_{j\in\mathbb{N}} \in \Lambda$ for all $\boldsymbol{\mu} \leq \boldsymbol{\nu}$. Here the ordering "$\leq$" on $\mathcal{F}$ is defined as $\mu_j \leq \nu_j$ for all $j \in \mathbb{N}$. We write $|\Lambda|$ to denote the finite cardinality of a set $\Lambda$. For $0 < p < \infty$, denote by $\ell^p(\mathcal{F})$ the space of

1 Index sets with the "downward closed" property are also referred to in the literature [57] as *lower sets*.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

sequences $\mathbf{t} = (t_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \subset \mathbb{R}$ satisfying $\|\mathbf{t}\|_{\ell^p(\mathcal{F})} := (\sum_{\boldsymbol{\nu}\in\mathcal{F}} |t_{\boldsymbol{\nu}}|^p)^{1/p} < \infty$. As usual, $\ell^\infty(\mathcal{F})$ equipped with the norm $\|\mathbf{t}\|_{\ell^\infty(\mathcal{F})} := \sup_{\boldsymbol{\nu}\in\mathcal{F}} |t_{\boldsymbol{\nu}}| < \infty$ denotes the space of all uniformly bounded sequences.

We consider the set $\mathbb{C}^{\mathbb{N}}$ endowed with the product topology. Any subset such as $[-1,1]^{\mathbb{N}}$ is understood to be equipped with the subspace topology. For $\varepsilon \in (0,\infty)$, we write $B_\varepsilon := \{z \in \mathbb{C} : |z| < \varepsilon\}$. Furthermore, $B_\varepsilon^{\mathbb{N}} := \times_{j\in\mathbb{N}} B_\varepsilon \subset \mathbb{C}^{\mathbb{N}}$. Elements of $\mathbb{C}^{\mathbb{N}}$ will be denoted by boldface characters such as $\boldsymbol{y} = (y_j)_{j\in\mathbb{N}} \in [-1,1]^{\mathbb{N}}$. For $\boldsymbol{\nu} \in \mathcal{F}$, the standard notations $\boldsymbol{y}^{\boldsymbol{\nu}} := \prod_{j\in\mathbb{N}} y_j^{\nu_j}$ and $\boldsymbol{\nu}! = \prod_{j\in\mathbb{N}} \nu_j!$ will be employed (throughout, $0! := 1$ and $0^0 := 1$, so that $\boldsymbol{\nu}!$ contains finitely many nontrivial factors). For any index set $\Lambda \subset \mathcal{F}$, we denote $\mathbb{P}_\Lambda := \operatorname{span}\{\boldsymbol{y}^{\boldsymbol{\nu}}\}_{\boldsymbol{\nu}\in\Lambda}$.

For a Banach space $X$, we denote by $\mathcal{P}(X)$ the space of Borel probability measures on $X$ and by $d_H(\cdot,\cdot)$ the Hellinger metric on $\mathcal{P}(X)$.

### 15.1.4 Structure of the present paper

The structure of this paper is as follows. In Section 15.2, we review the mathematical setting of Bayesian inverse problems for PDEs, including results that account for the impact of the PDE discretization error on the Bayesian posterior. In Section 15.3, we recall the notion of $(\boldsymbol{b},\varepsilon)$-holomorphic functions on polyellipses, taking values in Banach spaces and review approximation rate bounds for their truncated gPC expansion.

Sections 15.4–15.5 contain the mathematical core and main technical contributions of this paper: we define the DNN architectures and present, after recapitulating the basic operations of DNN calculus, expression rate bounds for so-called $(\boldsymbol{b},\varepsilon,\mathbb{R})$-holomorphic functions. This function class consists of maps from $[-1,1]^{\mathbb{N}}$ to $\mathbb{R}$, which allow holomorphic extensions (in each variable) to certain subsets of $\mathbb{C}^{\mathbb{N}}$. This is subsequently generalized to $(\boldsymbol{b},\varepsilon,\mathcal{X})$-holomorphic functions. To keep the network size possibly small, we employ a multilevel strategy by combining approximations to elements in $\mathcal{X}$ at different accuracy levels. Section 15.5.2 presents an illustrative example of a PDE with uncertain input data, which satisfy the preceding abstract hypotheses. Following this, we apply our result for $(\boldsymbol{b},\varepsilon,\mathbb{R})$-holomorphic functions to the Bayesian posterior density in Section 15.6. We show, in particular, that ReLU DNNs are able to express the posterior density with rates (in terms of the size of the DNN) free from the curse of dimensionality. We also show in Section 15.6.2 that DNNs allow for expression rates that are robust with respect to certain types of posterior concentration in the small noise and large data limits. Section 15.6.3 shows that the $L^\infty$-convergence of approximations of the posterior density implies the convergence of the approximate posterior measure in the Hellinger and total-variation distances. In Section 15.7, we give conclusions and indicate further directions. We provide proofs of all results from the main text in the extended version [61] of this work.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

## 15.2 Bayesian inverse UQ

We first present the abstract setting of BIP on function spaces, [24, 72, 76]. We then verify the abstract hypotheses in several examples, in particular, for diffusion equations with uncertain coefficients in polygons.

### 15.2.1 Forward model

We consider abstract parametric operator equations, possibly nonlinear, whose operators depend on uncertain input data $a$.

We consider given an uncertain input datum $a \in \tilde{X} \subset X$, where $X$ denotes a Banach space containing the set $\tilde{X}$ of admissible input data of the operator equation. Generally, $a$ is not accessible a priori and therefore is considered as uncertain input data. A priori knowledge about the distribution of $a \in X$ for a particular application is encoded through a probability measure $\mu_0$ on $X$, the Bayesian prior, which is supported on a measurable subset $\tilde{X} \subset X$ of admissible uncertain inputs. This implies, in particular, that $\tilde{X} \in \mathcal{B}(X)$ is $\mu_0$-measurable and that $\mu_0(\tilde{X}) = 1$; we discuss this in detail in Section 15.2.2.

The *abstract forward model* to be further considered given (a realization of) the uncertain input parameter $a \in \tilde{X}$ and a possibly nonlinear map $\mathcal{N}(a, \cdot) : \mathcal{X} \to \mathcal{Y}'$ is as follows:

$$\text{find } u \in \mathcal{X} : \quad \langle \mathcal{N}(a, u), v \rangle = 0 \quad \text{for all } v \in \mathcal{Y}. \tag{15.1}$$

Here $\langle \cdot, \cdot \rangle$ denotes the $\mathcal{Y}' \times \mathcal{Y}$ duality pairing. Throughout, we admit infinite-dimensional Banach spaces $X$, $\mathcal{X}$, $\mathcal{Y}$ (all results apply verbatim for the finite-dimensional settings).

In (15.1) the nonlinear map $\mathcal{N}(\cdot, \cdot) : X \times \mathcal{X} \to \mathcal{Y}'$ can be thought of as residual map for a PDE with solution space $\mathcal{X}$ and uncertain distributed input data $a$ from a function space $X$.

### 15.2.2 Bayesian inverse problem

We recapitulate the abstract setting of Bayesian inverse problems (BIPs) where the data-to-prediction map is constrained by possibly nonlinear operator equations (15.1), which are subject to unknown/unobservable input data.

#### 15.2.2.1 Setup

In the Bayesian inversion of the forward model (15.1), we in general have no access to the uncertain input $a$. Instead, we assume given *noisy observation data* $\delta \in Y$, where $Y$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

is a space of observation data. The data $\delta \in Y$ is a response of (15.1) for some admissible input $a \in \tilde{X}$ with response corrupted by additive observation noise $\eta \in Y$, i. e.,

$$\delta = \mathcal{G}(a) + \eta. \tag{15.2}$$

The *data-to-observation map* $\mathcal{G}(\cdot)$ is composed of the solution operator $G : a \mapsto u$ associated with (15.1) and a continuous linear observation map $\mathcal{O} \in \mathcal{L}(\mathcal{X}, Y)$ taking the solution $u(a) \in \mathcal{X}$ with input $a \in X$ to observations $\mathcal{O}(u(a)) \in Y$. Thus $\mathcal{G} : X \to Y : a \mapsto \mathcal{G}(a) := (\mathcal{O} \circ G)(a)$.

We often wish to predict a so-called *quantity of interest (QoI)*. In this work, we assume the QoI to be a bounded linear functional $Q \in \mathcal{L}(\mathcal{X}, Z)$, where $Z$ is a suitable Banach space. Then in this setup the inverse problem consists in estimating the "most likely" realization of the QoI based on solutions $u = G(a)$ of the forward problem (15.1), given noisy observation data $\delta$ of responses $\mathcal{G}(a)$.

In Bayesian inversion, we assume given a probability measure $\mu_0$ on the Banach space $X$ of inputs, which charges the set $\tilde{X} \subset X$ of admissible inputs, and which encodes our prior information about the occurrence of inputs $a$.

Given a realization of the parameter $a \in \tilde{X}$ and observation data $\delta \in Y$, we denote by $\mu^{\delta|a}$ the probability measure on $\delta$ conditioned on $a$. Under the *assumption* that $\mu^{\delta|a} \ll \mu_{\text{ref}}$ for some reference measure $\mu_{\text{ref}}$ on $Y$ and that $\mu^{\delta|a}$ has a density with respect to $\mu_{\text{ref}}$ that is positive $\mu_{\text{ref}}$-a. e., we may define the *likelihood potential* $\Phi(a; \delta) : X \times Y \to \mathbb{R}$ (the "negative log-likelihood") so that

$$\frac{\mathrm{d}\mu^{\delta|a}}{\mathrm{d}\mu_{\text{ref}}}(\delta) = \exp(-\Phi(a; \delta)), \quad \int_Y \exp(-\Phi(a; \delta))\mathrm{d}\mu_{\text{ref}}(\delta) = 1. \tag{15.3}$$

**Remark 15.2.1.** Let $Y = \mathbb{R}^K$ and assume that the observation noise $\eta \sim \mathcal{N}(0, \Gamma)$ is additive centered Gaussian with positive definite covariance matrix $\Gamma \in \mathbb{R}^{K \times K}$.

Then there exists a measure $\mu_{\text{ref}}$ on $Y$, equal to a $\Gamma$-dependent constant times the Lebesgue measure on $Y$, such that

$$\Phi(a; \delta) = \frac{1}{2}\left\|\Gamma^{-1/2}(\mathcal{G}(a) - \delta)\right\|_2^2 =: \frac{1}{2}\left\|\mathcal{G}(a) - \delta\right\|_\Gamma^2. \tag{15.4}$$

The potential $\Phi$ is an inverse covariance weighted, least squares functional of the response-to-observation misfit for uncertain input parameter $a \in X$ and observation data $\delta \in Y$.

In finite dimensions, Bayes' rule states that the posterior $\mu^{a|\delta}$ (the probability measure of the unknown $a$ conditioned on the data $\delta$) is proportional to the product of the likelihood $\mu^{\delta|a}$ and the prior $\mu_0$. In the present Banach space setting, this *formally* ex-

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

tends to

$$\frac{\mathrm{d}\mu^{a|\delta}}{\mathrm{d}\mu_0}(a) = \frac{1}{Z(\delta)}\exp(-\Phi(a;\delta)), \quad \text{where } Z(\delta) = \int_X \exp(-\Phi(a;\delta))\mathrm{d}\mu_0(a), \tag{15.5}$$

which can be made rigorous; see the references below. Here $Z(\delta)$ is a normalization constant guaranteeing $\frac{\exp(-\Phi(a;\delta))}{Z(\delta)}$ to be a probability density as a function of $a \in X$ with respect to the measure $\mu_0$. In the Bayesian methodology the posterior probability measure $\mu^{a|\delta}$ is considered an updated version of the prior $\mu_0$ on the uncertain inputs that is informed by the observation data $\delta$. In the following, we denote the posterior probability measure by $\mu^{\delta}$.

**Remark 15.2.2.** Note that (15.5) is independent of the choice of reference measure $\mu_{\mathrm{ref}}$ in (15.3): Let $\tilde{\mu}_{\mathrm{ref}}$ be another (equivalent) reference measure such that $\frac{\mathrm{d}\mu^{\delta|a}}{\mathrm{d}\tilde{\mu}_{\mathrm{ref}}}(\delta) = \exp(-\tilde{\Phi}(a;\delta))$. Then

$$\exp(-\Phi(a;\delta)) = \frac{\mathrm{d}\mu^{\delta|a}}{\mathrm{d}\mu_{\mathrm{ref}}}(\delta) = \frac{\mathrm{d}\mu^{\delta|a}}{\mathrm{d}\tilde{\mu}_{\mathrm{ref}}}(\delta)\frac{\mathrm{d}\tilde{\mu}_{\mathrm{ref}}}{\mathrm{d}\mu_{\mathrm{ref}}}(\delta) = \exp(-\tilde{\Phi}(a;\delta))\frac{\mathrm{d}\tilde{\mu}_{\mathrm{ref}}}{\mathrm{d}\mu_{\mathrm{ref}}}(\delta).$$

Hence $\tilde{\Phi}(a;\delta) = \Phi(a;\delta) + \log(c)$ with the $a$-independent constant $c = \frac{\mathrm{d}\tilde{\mu}_{\mathrm{ref}}}{\mathrm{d}\mu_{\mathrm{ref}}}(\delta)$. The constant $c$ will merely influence the normalization $Z(\delta)$ in (15.5), but either choice of $\mu_{\mathrm{ref}}$ or $\tilde{\mu}_{\mathrm{ref}}$ leads to the same formula for $a \mapsto \frac{\mathrm{d}\mu^{a|\delta}}{\mathrm{d}\mu_0}(a)$.

We refer to [21, 39, 76] for a detailed discussion and further references, in particular, [21, Section 3.4.2], which is an application of the more general discussion in [21, Section 3.2]. Our definition of the likelihood potential $\Phi$ in (15.4) is consistent with [21, Equation (10.39)]. It is shifted with respect to the definition in [21, Equation (10.30)] by adding to $\Phi$ a function depending on $\delta$, but not on $a$; see Remark 15.2.2 and also [21, Remark 5].

#### 15.2.2.2 Assumptions

Based on [20, 21, 39, 76], we now formalize the preceding concepts. To this end, we introduce a set of assumptions on the prior and on the forward map, which ensure the well-posedness and continuous dependence of the BIP.

**Assumption 15.2.3** ([39, Assumption 2.1]). In the Banach space $X$ of uncertain parameters and the Banach space $Y$ of observation data, the potential $\Phi : X \times Y \to \mathbb{R}$ satisfies:

(i) (bounded below) There is $\alpha_1 \geq 0$ such that for every $r > 0$, there exists a constant $M(\alpha_1, r) \in \mathbb{R}$ such that for every $u \in X$ and every data $\delta \in Y$ with $\|\delta\|_Y < r$, we have

$$\Phi(u;\delta) \geq M - \alpha_1\|u\|_X.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

(ii) (boundedness above) For every $r > 0$, there exists $K(r) > 0$ such that for every $u \in X$ and every $\delta \in Y$ with $\max\{\|u\|_X, \|\delta\|_Y\} < r$, we have

$$\Phi(u;\delta) \leq K.$$

(iii) (Lipschitz continuous dependence on $u$) For every $r > 0$, there exists a constant $L(r) > 0$ such that for every $u_1, u_2 \in X$ and every $\delta \in Y$ with $\max\{\|u_1\|_X, \|u_2\|_X, \|\delta\|_Y\} < r$, we have

$$\left|\Phi(u_1;\delta) - \Phi(u_2;\delta)\right| \leq L\|u_1 - u_2\|_X.$$

(iv) (Lipschitz continuity with respect to observation data $\delta \in Y$) For some $\alpha_2 \geq 0$ and for every $r > 0$, there exists $C(\alpha_2, r) \in \mathbb{R}$ such that for every $\delta_1, \delta_2 \in Y$ with $\max\{\|\delta_1\|_Y, \|\delta_2\|_Y\} < r$ and every $u \in X$, we have

$$\left|\Phi(u;\delta_1) - \Phi(u;\delta_2)\right| \leq \exp(\alpha_2\|u\|_X + C)\|\delta_1 - \delta_2\|_Y.$$

(v) (Radon prior measure) The prior measure $\mu_0$ is a Radon probability measure charging a measurable subset $\tilde{X} \subseteq X$ with $\tilde{X} \in \mathcal{B}(X)$ of admissible uncertain parameters, i. e., $\mu_0(\tilde{X}) = 1$.

(vi) (exponential tails) The prior measure $\mu_0$ on the Banach space $X$ has exponential tails:

$$\exists \kappa > 0 : \quad \int_X \exp(\kappa\|u\|_X)\mathrm{d}\mu_0(u) < \infty. \tag{15.6}$$

**Remark 15.2.4.** Assumption (v) on the prior $\mu_0$ being a Radon probability measure is always satisfied when $X$ is separable.

#### 15.2.2.3 Well-posedness

We consider the well-posedness of the BIP in the following sense.

**Definition 15.2.5** (Well-posedness of the BIP, [39, Definition 1.4]). For Banach spaces $X$, $Y$ with $d_H(\cdot,\cdot)$ denoting the Hellinger metric on the space $\mathcal{P}(X)$ of Borel probability measures on $X$, for a prior $\mu_0 \in \mathcal{P}(X)$ and for the likelihood potential $\Phi$, the BIP (15.5) is well posed if the following holds:

(i) (existence and uniqueness) For every data $\delta \in Y$, there exists a unique posterior measure $\mu^\delta \in \mathcal{P}(X)$, which is absolutely continuous with respect to the prior $\mu_0$, and which satisfies (15.5).

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

(ii) (stability) For all $\varepsilon > 0$ and $r > 0$, there exists a constant $C_\varepsilon(r) > 0$ such that for every $\delta, \delta' \in Y$ with $\max\{\|\delta\|_Y, \|\delta'\|_Y\} < r$ and $\|\delta - \delta'\|_Y \leq C_\varepsilon$, we have

$$d_H(\mu^\delta, \mu^{\delta'}) < \varepsilon.$$

#### 15.2.2.4 Existence and continuous dependence

We are now in position to state sufficient conditions for the well-posedness of the BIP and for the existence and uniqueness of the posterior $\mu^\delta$. We work in the abstract setting Assumption 15.2.3, deferring the verification of the items in Assumption 15.2.3 to the ensuing discussion of concrete model problems.

**Theorem 15.2.6** ([39, Theorems 2.4 and 2.6]). *Let $X$ and $Y$ be Banach spaces, and let a likelihood function $\Phi : X \times Y \to \mathbb{R}$ satisfy Assumption* 15.2.3, *items* (i), (ii), (iii) *with some $\alpha_1 > 0$. Moreover, suppose the prior measure $\mu_0 \in \mathcal{P}(X)$ satisfies Assumption* 15.2.3, *items* (v) *and* (vi), *with some constant $\kappa > 0$.*

*Then:*

(i) *If $\kappa \geq \alpha_1$, for every $\delta \in Y$, the posterior measure $\mu^\delta$ defined in Equation* (15.5) *is a well-defined Radon probability measure on $X$.*

(ii) *(Lipschitz continuity of the posterior with respect to the data) If $\Phi$ satisfies in addition Assumption* 15.2.3, *item* (iv), *with some constant $\alpha_2 \geq 0$ and if the constant $\kappa$ from Assumption* 15.2.3, *item* (vi), *satisfies $\kappa \geq \alpha_1 + 2\alpha_2$, then for every $r > 0$, there exists a constant $C(r) > 0$ such that, for all $\delta, \delta' \in Y$ with $\max\{\|\delta\|_Y, \|\delta'\|_Y\} < r$, the posteriors $\mu^\delta, \mu^{\delta'} \in \mathcal{P}(X)$ satisfy*

$$d_H(\mu^\delta, \mu^{\delta'}) \leq C(r)\|\delta - \delta'\|_Y. \tag{15.7}$$

A proof of this result is, for example, in [39, Theorems 2.4 and 2.6].

#### 15.2.2.5 Consistent approximation

In the numerical approximation of posteriors $\mu^\delta$ where the input-to-observation map $\mathcal{G} = \mathcal{O} \circ G : X \to Y$ involves a well-posed parametric forward operator equation (15.1), we will in general have to resort to approximate numerical solutions of (15.1). Generically, we tag such approximate solution maps by a subscript $N \in \mathbb{N}$, which should be understood as the "number of degrees of freedom" involved in the discretization of the parametric equation (15.1). In this way, we denote the data-to-solution map of the nonlinear equation (15.1) by $G_N : X \to \mathcal{X}$, the corresponding data-to-observation map by $\mathcal{G}_N = \mathcal{O} \circ G_N$, and the likelihood potential by $\Phi_N$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Approximation of the forward model (15.1), e. g., by consistent discretization, leads to an *approximate Bayesian inverse problem* of the form

$$\frac{d\mu_N^\delta}{d\mu_0}(a) = \frac{1}{Z_N(\delta)} \exp(-\Phi_N(a;\delta)), \quad \text{where } Z_N(\delta) := \int_X \exp(-\Phi_N(a;\delta))\mathrm{d}\mu_0(a). \tag{15.8}$$

Assuming exact observations $\mathcal{O}(\cdot)$ at hand, the approximate potential $\Phi_N$ in (15.8) is

$$\Phi_N(a;\delta) = \frac{1}{2}\|\Gamma^{-1/2}((\mathcal{O}\circ G_N)(a) - \delta)\|_2^2, \quad a \in \tilde{X}, \delta \in Y.$$

The posterior $\mu^\delta$ would, consequently, also be approximated by the corresponding numerical posterior, which we denote by $\mu_N^\delta$.

It is of interest to identify sufficient conditions so that, as $N \to \infty$, the approximate posteriors $\{\mu_N^\delta\}_{N\geq 1}$ tend to the posterior $\mu^\delta$ in $\mathcal{P}(X)$.

**Definition 15.2.7** (Consistent posterior approximation, [39, Definition 1.5]). The approximate Bayesian inverse problem (15.8) is said to be a *consistent approximation of* (15.5) *for a prior* $\mu_0 \in \mathcal{P}(X)$ *and a potential* $\Phi$ if the approximate potential $\Phi_N$ is such that for every data $\delta \in Y$, as $N \to \infty$, we have that

$$|\Phi(a;\delta) - \Phi_N(a;\delta)| \to 0 \quad \text{implies} \quad d_H(\mu^\delta, \mu_N^\delta) \to 0.$$

Apart from consistency in the sense of Definition 15.2.7, in the numerical approximation of BIPs, we are also interested in *convergence rates*: if the numerical approximation $G_N$ of the forward solution map converges with a certain rate, say $\psi(N)$, with $\psi$ a nonnegative function such that $\psi(N) \downarrow 0$ as $N \to \infty$, then the corresponding posteriors $\mu_N^\delta$ should converge with a rate related to $\psi(N)$. The following theorem, which is proved in [21, Theorem 18], gives sufficient conditions for the posterior convergence.

**Theorem 15.2.8** ([21, Theorem 18]). *Let $X$ and $Y$ be given Banach spaces of uncertain parameters $a$ and observation data $\delta$, respectively.*

*Let $\mu_0 \in \mathcal{P}(X)$ be a Borel probability measure on $X$ satisfying Assumption* 15.2.3, *items* (i)–(vi), *so that for observation data $\delta \in Y$, the BIPs* (15.5) *and* (15.8) *for $\mu^\delta, \mu_N^\delta \in \mathcal{P}(X)$ are well-defined.*

*Assume also that the likelihood potentials $\Phi$ and $\Phi_N$ satisfy Assumption* 15.2.3, *items* (i), (ii) *with constant $\alpha_1 \geq 0$, which is uniform with respect to $N$, and that for some $\alpha_3 \geq 0$, there exists $C(\alpha_3) > 0$ independent of $N$ such that for every $a \in \tilde{X}$,*

$$|\Phi(a;\delta) - \Phi_N(a;\delta)| \leq C\exp(\alpha_3\|a\|_X)\psi(N) \tag{15.9}$$

*with $\psi(N) \downarrow 0$ as $N \to \infty$.*

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*If furthermore Assumption* 15.2.3, *item* (vi), *holds with* $\kappa \geq \alpha_1 + 2\alpha_3$, *then for every* $r > 0$, *there exists a constant* $D(r) > 0$ *such that for every* $\delta \in Y$ *with* $\|\delta\|_Y < r$,

$$\forall N \in \mathbb{N}: \quad d_H(\mu^\delta, \mu_N^\delta) \leq D\psi(N).$$

*Here the constant* $D(r)$ *generally depends on the covariance* $\Gamma$ *of the centered Gaussian observation noise* $\eta$ *in* (15.2).

## 15.2.3 Prior modeling

The modeling of prior probability measures on function spaces of distributed uncertain PDE input data $a$ in the model (15.1) has been developed in several references in recent years. The "usual construction" is based on (a) coordinate representations of (realizations of) instances of $a$ in terms of a suitable basis $\{\psi_j\}_{j\geq 1}$ (thereby implying $a$ will take values in a separable subset $\bar{X}$ of $X$) and on (b) construction of the prior as a countable product probability measure of probability measures on the coordinate spaces.

This approach, which is inspired by N. Wiener's construction of the Wiener process by placing Gaussian measures on coefficient realizations of Fourier series, has been realized, for example, in [20, 33, 44] for Besov spaces, and in [39, 77] and the references therein for more general priors.

## 15.2.4 Examples

The foregoing abstract setting (15.1) accommodates a wide range of PDE boundary value, eigenvalue, control, and shape optimization problems with uncertain function space input $a \in X$. We illustrate the scope by listing several examples, which are covered by the ensuing abstract DNN expression rate bounds. In all examples, $\mathrm{D} \subset \mathbb{R}^d$ will denote an open, bounded, connected, and polytopal domain in physical Euclidean space of dimension $d \geq 2$. In dimension $d = 1$, D will denote an open bounded interval of positive length.

### 15.2.4.1 Diffusion equation

We consider a linear second-order diffusion equation with uncertain coefficients in $\mathrm{D} \subset \mathbb{R}^2$. Holomorphic dependence of solutions on coefficient data was shown in [5], and the numerical analysis, including finite-element discretization in D on corner-refined families of triangulations, with approximation rate estimates for both, the parametric solution, and the Karhunen–Loeve expansion terms, was provided in [34]. Given a source term $f \in H^{-1}(\mathrm{D}) = (H_0^1(\mathrm{D}))^*$ and an isotropic diffusion coefficient

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$a \in \tilde{X} \subset \{a \in L^\infty(\mathrm{D}) : \operatorname{ess\,inf}_{\boldsymbol{x}\in\mathrm{D}} a(\boldsymbol{x}) > 0\}$, the diffusion problem reads as follows: find $u \in H^1_0(\mathrm{D})$ such that

$$\mathcal{N}(a,u)(\boldsymbol{x}) := f(\boldsymbol{x}) + \nabla\cdot\big(a(\boldsymbol{x})\nabla u(\boldsymbol{x})\big) = 0 \quad \text{in D}, \quad u|_{\partial\mathrm{D}} = 0. \tag{15.10}$$

It falls into the variational setting (15.1) with $\mathcal{X} = \mathcal{Y} = H^1_0(\mathrm{D})$ and $X = L^\infty(\mathrm{D})$. In [5, 34], also anisotropic diffusion coefficients $a$ and advection and reaction terms were admitted.

For $a \in \tilde{X}$, the weak formulation (15.1) of (15.10) is uniquely solvable, and the data-to-solution map $G : \tilde{X} \to \mathcal{X} : a \mapsto u(a)$ is continuous. Equipping $\mathcal{X}$ with the norm $\|v\|_{\mathcal{X}} = \|\nabla v\|_{L^2(\mathrm{D})}$, we have

$$\|u\|_{\mathcal{X}} \le \frac{\|f\|_{H^{-1}(\mathrm{D})}}{\operatorname{ess\,inf}_{\boldsymbol{x}\in\mathrm{D}} a(\boldsymbol{x})}.$$

Assuming *affine-parametric* uncertain input [15, 16, 73], i. e., given $a_0 \in X$ with

$$a_- := \operatorname*{ess\,inf}_{\boldsymbol{x}\in\mathrm{D}} a_0(\boldsymbol{x}) > 0$$

for $\{\psi_j\}_{j\ge1} \subset X$ with $\sum_{j\ge1}\|\psi_j\|_X < a_-$, we choose the prior such that its support is contained in the set

$$\tilde{X} := \left\{a \in X : a = a(\boldsymbol{y}) := a_0 + \sum_{j\ge1} y_j\psi_j,\ \boldsymbol{y} = (y_j)_{j\ge1} \in [-1,1]^{\mathbb{N}}\right\}. \tag{15.11}$$

For every $\boldsymbol{y} \in [-1,1]^{\mathbb{N}}$ and $a(\boldsymbol{y}) \in \tilde{X}$, problem (15.10) admits a unique parametric solution $u(\boldsymbol{y}) \in \mathcal{X}$ such that $\mathcal{N}(a(\boldsymbol{y}), u(\boldsymbol{y})) = 0$ in $H^{-1}(\mathrm{D})$.

#### 15.2.4.2 Elliptic eigenvalue problem with uncertain coefficient

For $a \in \tilde{X}$ as defined in (15.11), for every $\boldsymbol{y} \in [-1,1]^{\mathbb{N}}$, we seek solutions $(\lambda(\boldsymbol{y}), w(\boldsymbol{y})) \in \mathbb{R}\times H^1_0(\mathrm{D})\backslash\{0\}$ of the eigenvalue problem

$$\mathcal{N}\big(a(\boldsymbol{y}), (\lambda(\boldsymbol{y}), w(\boldsymbol{y}))\big) = 0 \quad \text{in } H^{-1}(\mathrm{D}), \tag{15.12}$$

where, for every $a \in \tilde{X}$, $\mathcal{N}(a,(\lambda,w)) : \mathbb{R}\times H^1_0(\mathrm{D}) \to H^{-1}(\mathrm{D}) : (\lambda, w) \mapsto \lambda w + \nabla\cdot(a\nabla w)$. For every $\boldsymbol{y}$, the EVP (15.12) admits a sequence $\{(\lambda_k(\boldsymbol{y}), w_k(\boldsymbol{y})) : k = 1,2,\dots\}$ of real eigenvalues $\lambda_k(\boldsymbol{y})$ (which we assume enumerated according to their size with counted multiplicity) with associated eigenfunctions $w_k(\boldsymbol{y}) \in H^1_0(\mathrm{D})$ (which form a dense set in $H^1_0(\mathrm{D})$). It is known (e. g., [28, Proposition 2.4]) that the first eigenpair $\{(\lambda_1(\boldsymbol{y}), w_1(\boldsymbol{y})) : \boldsymbol{y} \in [-1,1]^{\mathbb{N}}\}$ is isolated and admits a uniform (with respect to $\boldsymbol{y} \in [-1,1]^{\mathbb{N}}$) spectral gap.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

# 15.3 Generalized polynomial chaos surrogates

## 15.3.1 Uncertainty parametrization

Let $\mathcal{Z}$ and $\mathcal{X}$ be two complex Banach spaces, and let $(\psi_j)_{j\in\mathbb{N}}$ be a sequence in $\mathcal{Z}$. Additionally, suppose that $O \subseteq \mathcal{Z}$ is open, and let $\mathfrak{u} : O \to \mathcal{X}$ be complex differentiable. With the parameter domain $U := [-1,1]^{\mathbb{N}}$, we consider the infinite parametric map

$$u(\boldsymbol{y}) := \mathfrak{u}\Big(\sum_{j\in\mathbb{N}} y_j\psi_j\Big) \quad \forall \boldsymbol{y} = (y_j)_{j\in\mathbb{N}} \in U, \tag{15.13}$$

which is well-defined, for instance, if $(\|\psi_j\|_{\mathcal{Z}})_{j\in\mathbb{N}} \in \ell^1(\mathbb{N})$. Here the map $U \to O : \boldsymbol{y} \mapsto \sum_{j\in\mathbb{N}} y_j\psi_j$ is understood as an (affine) parameterization of the uncertain input $a$, and $\mathfrak{u}$ denotes the map that relates the input to the solution of the model under consideration.

Under certain assumptions, such maps allow a representation as a sparse *Taylor generalized polynomial chaos expansion* [15, 16], i. e., for $\boldsymbol{y} \in U$,

$$u(\boldsymbol{y}) = \sum_{\boldsymbol{\nu}\in\mathcal{F}} t_{\boldsymbol{\nu}}\boldsymbol{y}^{\boldsymbol{\nu}}, \quad t_{\boldsymbol{\nu}} = \frac{1}{\boldsymbol{\nu}!}\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}u(\boldsymbol{y})|_{\boldsymbol{y}=\boldsymbol{0}} \in \mathcal{X}, \tag{15.14}$$

or as a sparse *Legendre generalized polynomial chaos expansion* [13], i. e.,

$$u(\boldsymbol{y}) = \sum_{\boldsymbol{\nu}\in\mathcal{F}} l_{\boldsymbol{\nu}}L_{\boldsymbol{\nu}}(\boldsymbol{y}), \quad l_{\boldsymbol{\nu}} = \int_U L_{\boldsymbol{\nu}}(\boldsymbol{y})u(\boldsymbol{y})\mathrm{d}\mu_U(\boldsymbol{y}) \in \mathcal{X}, \tag{15.15}$$

where $L_{\boldsymbol{\nu}}(\boldsymbol{y}) = \prod_{j\in\mathbb{N}} L_{\nu_j}(y_j)$, and $L_n : [-1,1] \to \mathbb{R}$ denotes the $n$th Legendre polynomial normalized in $L^2([-1,1], \lambda/2)$, where $\lambda$ denotes the Lebesgue measure on $[-1,1]$, i. e., $\lambda/2$ is a uniform probability measure on $[-1,1]$. Also, $\mu_U := \bigotimes_{j\in\mathbb{N}} \frac{\lambda}{2}$ denotes the uniform probability measure on $U = [-1,1]^{\mathbb{N}}$ equipped with the product $\sigma$-algebra. Then by [59, § 18.3]

$$\|L_n\|_{L^\infty([-1,1])} \leq (1+2n)^{\frac{1}{2}} \quad \forall n \in \mathbb{N}_0. \tag{15.16}$$

The summability properties of the ($\mathcal{X}$-norms of) Taylor or Legendre gPC coefficients $(\|t_{\boldsymbol{\nu}}\|_{\mathcal{X}})_{\boldsymbol{\nu}\in\mathcal{F}}$, $(\|l_{\boldsymbol{\nu}}\|_{\mathcal{X}})_{\boldsymbol{\nu}\in\mathcal{F}}$ are key for assigning a meaning to such formal gPC expansions like (15.14) and (15.15). For example, as for every $\boldsymbol{y} \in U$ and every $\boldsymbol{\nu} \in \mathcal{F}$ we have $|\boldsymbol{y}^{\boldsymbol{\nu}}| \leq 1$, the summability $(\|t_{\boldsymbol{\nu}}\|_{\mathcal{X}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^1(\mathcal{F})$ guarantees the unconditional convergence in $\mathcal{X}$ of the series in (15.14) for every $\boldsymbol{y} \in U$. As we will recall in Section 15.3.3, this summability is in turn ensured by a suitable form of holomorphic continuation of the parameter-to-response map $u : U \to \mathcal{X}$.

**Remark 15.3.1.** We assume here $\mathcal{X}$ to be a complex space. If $\mathcal{X}$ is a Banach space over $\mathbb{R}$, we can consider $\mathfrak{u}$ as a map to the complexification $\mathcal{X}_{\mathbb{C}} = \mathcal{X} + \mathrm{i}\mathcal{X}$ of $\mathcal{X}$ equipped with

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

the so-called *Taylor norm* $\|v + \mathrm{i}w\|_{\mathcal{X}_{\mathbb{C}}} := \sup_{t\in[0,2\pi)} \|\cos(t)v - \sin(t)w\|_{\mathcal{X}}$ for $v, w \in \mathcal{X}$ (cp. [56]). Here $\mathrm{i} = \sqrt{-1}$ with $\arg(\mathrm{i}) = \pi/2$.

## 15.3.2 $(\boldsymbol{b}, \boldsymbol{\varepsilon}, \mathcal{X})$-holomorphy

To prove expressive power estimates for DNNs, we use parametric holomorphic maps from a compact parameter domain $U$ into a Banach space $\mathcal{X}$ with quantified sizes of domains of holomorphy. To introduce such maps, we recapitulate principal definitions and results from [12, 13, 16, 83] and the references therein. The notion of $(\boldsymbol{b},\varepsilon)$-holomorphy (given in Definition 15.3.3), which stipulates holomorphic parameter dependence of a function $u : U \to \mathcal{X}$ in each variable on certain product domains $\mathcal{O} = \times_{j\in\mathbb{N}} O_j \subseteq \mathbb{C}^{\mathbb{N}}$, has been found to be a sufficient condition on a parametric map $U \ni \boldsymbol{y} \mapsto u(\boldsymbol{y}) \in \mathcal{X}$, in order that $u$ admits gPC expansions with $p$-summable coefficients for some $p \in (0,1)$; see, e. g., [13, 74] and also Section 15.3.3. In the following, we extend the results from [74] in the sense that we admit smaller domains of holomorphy: each $O_j = \mathcal{E}_{\rho_j}$ is a *Bernstein-ellipse* defined by

$$\mathcal{E}_\rho := \left\{ \frac{z + z^{-1}}{2} \,:\, z \in \mathbb{C},\ 1 \le |z| < \rho \right\} \subseteq \mathbb{C},$$

rather than a complex disc $O_j = B_{\rho_j}$ as in [74].

**Remark 15.3.2.** Let $\mathcal{J} \subseteq \mathbb{N}$. Throughout, the *continuity of a function defined on a cylindrical set* $\times_{j\in\mathcal{J}} O_j$ with $O_j \subseteq \mathbb{C}$ for all $j \in \mathcal{J}$ will be understood as the continuity with respect to the subspace topology on $\times_{j\in\mathcal{J}} O_j \subset \times_{j\in\mathcal{J}} \mathbb{C}$, where $\times_{j\in\mathcal{J}} \mathbb{C}$ is assumed to be equipped with the product topology by our convention (see Section 15.1.3). In this topology the parameter domain $U = [-1,1]^{\mathbb{N}}$ is compact by Tikhonov's theorem [55, Theorem 37.3].

In the following, if $\boldsymbol{\rho} = (\rho_j)_{j=1}^N \subseteq (1,\infty)$ for some $N \in \mathbb{N}$, then we define the polyellipse $\mathcal{E}_{\boldsymbol{\rho}} := \times_{j=1}^N \mathcal{E}_{\rho_j} \subseteq \mathbb{C}^N$, and similarly in case $\boldsymbol{\rho} = (\rho_j)_{j\in\mathbb{N}} \subseteq (1,\infty)$,

$$\mathcal{E}_{\boldsymbol{\rho}} := \bigtimes_{j\ge1} \mathcal{E}_{\rho_j} \subseteq \mathbb{C}^{\mathbb{N}}.$$

**Definition 15.3.3** ($(\boldsymbol{b},\varepsilon,\mathcal{X})$-holomorphy). Let $\mathcal{X}$ be a complex Banach space. Let $\boldsymbol{b} = (b_j)_{j\in\mathbb{N}}$ be a decreasing sequence of positive reals $b_j$ such that $\boldsymbol{b} \in \ell^p(\mathbb{N})$ for some $p \in (0,1]$.

We say that a map $u : U \to \mathcal{X}$ is $(\boldsymbol{b},\varepsilon,\mathcal{X})$-holomorphic if there exists a constant $M < \infty$ such that

(i) $u : U \to \mathcal{X}$ is continuous,

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

(ii) for every sequence $\boldsymbol{\rho} = (\rho_j)_{j\in\mathbb{N}} \subset (1,\infty)^{\mathbb{N}}$ that is $(\boldsymbol{b},\varepsilon)$-*admissible*, i. e., satisfies

$$\sum_{j\in\mathbb{N}} b_j(\rho_j - 1) \leq \varepsilon, \tag{15.17}$$

$u$ admits a separately holomorphic extension (again denoted by $u$) onto the polyellipse $\mathcal{E}_{\boldsymbol{\rho}}$,

(iii) for each $(\boldsymbol{b},\varepsilon)$-admissible $\boldsymbol{\rho}$, we have

$$\sup_{\boldsymbol{z}\in\mathcal{E}_{\boldsymbol{\rho}}} \|u(\boldsymbol{z})\|_{\mathcal{X}} \leq M. \tag{15.18}$$

If it is clear from the context that $\mathcal{X} = \mathbb{C}$, then we will omit $\mathcal{X}$ in the notation.

**Remark 15.3.4.** We note that for $\boldsymbol{b} \in \ell^1(\mathbb{N})$ as in Definition 15.3.3, $b_j \to 0$ as $j \to \infty$. By (15.17), $(\boldsymbol{b},\varepsilon)$-admissible polyradii $\boldsymbol{\rho}$ can satisfy $\rho_j \to \infty$, implying that the component sets $\mathcal{E}_{\rho_j}$ grow as $j \to \infty$. We also observe the following elementary geometric fact:

$$\forall \rho > 1: \quad \mathcal{E}_\rho \supset B_{(\rho-1/\rho)/2}. \tag{15.19}$$

In particular, $\mathcal{E}_\rho \supset \overline{B_1} \supset [-1,1]$ for all $\rho > 1 + \sqrt{2}$. Bernstein ellipses $\mathcal{E}_\rho$ are moreover useful if the domain of holomorphy of $u$ does not contain $B_1$. Moreover, if $\rho_j \to \infty$, after all but a (possibly small) finite number of parameters, the domains of holomorphy $\mathcal{E}_{\rho_j}$ contain a polydisc of radius $(\rho_j - 1/\rho_j)/2 > 1$. We will see in Section 15.4 that multivariate monomials can be expressed by smaller DNNs than, e. g., multivariate Legendre or Jacobi polynomials. In particular, for the emulation of tensor products of Taylor monomials, the product network is of smaller size than that for the emulation of tensor product Legendre polynomials. The reason is that the $L^\infty$-norm of Taylor monomials equals 1, whereas for $\boldsymbol{\nu} \in \mathcal{F}$, we have $\|L_{\boldsymbol{\nu}}\|_{L^\infty(U)} \leq \prod_{j\in\operatorname{supp}\boldsymbol{\nu}} \sqrt{1+2\nu_j}$ (cf. (15.16)). Due to the growth of this bound, to achieve the same absolute accuracy, a larger relative accuracy and thus a larger product network size are required (see Proposition 15.4.3). We therefore use in our expression rate bounds "Taylor DNN emulations" as in [74] for all but a fixed finite number of dimensions. There we use an *exponential expression rate bound* from [62] for the ReLU DNN approximation of tensor product Legendre polynomials (Proposition 15.4.6).

Definition 15.3.3 has been similarly stated in [13]. The sequence $\boldsymbol{b}$ in Definition 15.3.3 quantifies the size of the domains of analytic continuation of the parametric map with respect to the parameters $y_j \in \boldsymbol{y}$: the stronger the decrease of $\boldsymbol{b}$, the faster the radii $\rho_j$ of $(\boldsymbol{b},\varepsilon)$-admissible sequences $\boldsymbol{\rho}$ increase. The sequence $\boldsymbol{b}$ (or, more precisely, the summability exponent $p$ such that $\boldsymbol{b} \in \ell^p(\mathbb{N})$) determines the algebraic rate at which the gPC coefficients tend to 0 (see Theorem 15.3.7). The notion

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

of $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphy applies to large classes of parametric operator equations, notably including functions of the type (15.13). This statement is given in the next lemma, which is proven in [81, Lemma 2.2.7]; see also [83, Lemma 3.3] (for a version based on holomorphy on polydiscs rather than on polyellipses).

**Lemma 15.3.5.** *Let $\mathfrak{u} : O \to \mathcal{X}$ be holomorphic, where $O \subseteq \mathcal{Z}$ is open. Assume that $(\psi_j)_{j\in\mathbb{N}} \subseteq \mathcal{Z}$, $\psi_j \neq 0$ for all $j$, with $(\|\psi_j\|_{\mathcal{Z}})_{j\in\mathbb{N}} \in \ell^1(\mathbb{N})$ and $\{\sum_{j\in\mathbb{N}} y_j\psi_j : \boldsymbol{y} \in U\} \subseteq O$. Then there exists $\varepsilon > 0$ such that $u(\boldsymbol{y}) = \mathfrak{u}(\sum_{j\in\mathbb{N}} y_j\psi_j)$, $\boldsymbol{y} \in U$ defines a $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphic function with $b_j := \|\psi_j\|_{\mathcal{Z}}$.*

### 15.3.3 Summability of gPC coefficients

As mentioned above, the relevance of $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphy lies in that it guarantees such functions to possess gPC expansions with coefficients whose norms are $p$-summable for some $p \in (0, 1)$. This $p$-summability is the crucial property required to establish the convergence rates of certain partial sums. Our analysis of the expressive power of DNNs of such parametric solution families will be based on a version of these results as stated in the next theorem. To reduce the asymptotic size of the networks, we consider gPC expansions combining both multivariate monomials and multivariate Legendre polynomials, as motivated in Remark 15.3.4. Whereas $p$-summability of the norms of both the Taylor and Legendre coefficients of such functions is well known (under suitable assumptions), Theorem 15.3.7 below is not available in the literature. For this reason, we provide a proof but stress that the general line of arguments closely follows earlier works such as [13, 15, 16, 82].

In the next theorem, we distinguish between low- and high-dimensional coordinates: We will use in "low dimensions" indexed by $j \in \{1, \dots, J\}$ Legendre expansions, whereas in the coordinates indexed by $j > J$, we resort to Taylor gPC expansions. For $1 \leq j \leq J$, we thus exploit holomorphy on polyellipses $\mathcal{E}_{\rho_j}$ and Legendre gPC expansions. For $j > J$, we emulate by ReLU DNNs the corresponding Taylor gPC expansions in these coordinates using [74] and the fact that sufficiently large Bernstein ellipses with foci $\pm 1$ contain discs with radius $> 1$ centered at the origin (as pointed out in Remark 15.3.4).

Accordingly, we introduce the following notation: for some fixed $J \in \mathbb{N}$ (defined in the following) and $\boldsymbol{\nu} \in \mathcal{F}$, set

$$\boldsymbol{\nu}_E := (\nu_1, \dots, \nu_J), \quad \boldsymbol{\nu}_F := (\nu_{J+1}, \nu_{J+2}, \dots),$$

and $\mathcal{F}_E := \mathbb{N}_0^J$, and we will write $\boldsymbol{\nu} = (\boldsymbol{\nu}_E, \boldsymbol{\nu}_F)$. Moreover, $U_E := [-1, 1]^J$ and $U_F := \times_{j>J}[-1, 1]$, and for $\boldsymbol{y} = (y_j)_{j\in\mathbb{N}} \in U$, define $\boldsymbol{y}_E := (y_j)_{j=1}^J \in U_E$ and $\boldsymbol{y}_F := (y_j)_{j>J} \in U_F$. In particular, we will employ the notation $\boldsymbol{y}_F^{\boldsymbol{\nu}_F} = \prod_{j>J} y_j^{\nu_j}$. Additionally, for a function $u : U \to \mathcal{X}$, by $u(\boldsymbol{y}_E, \boldsymbol{0})$ we mean $u$ evaluated at $(y_1, \dots, y_J, 0, 0, \dots) \in U$. In terms of the Lebesgue measure $\lambda$ on $[-1, 1]$, define $\mu_E := \bigotimes_{j=1}^J \frac{\lambda}{2}$ on $U_E$ and $\mu_F := \bigotimes_{j>J} \frac{\lambda}{2}$ on $U_F$.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Lemma 15.3.6.** *Let $C_0 := 4/9$. Then $B^{\mathbb{C}}_{C_0\rho} \subseteq \mathcal{E}_\rho$ for all $\rho \geq 3$.*

*Proof.* By Remark 15.3.4 we have $B_{(\rho-\rho^{-1})/2} \subseteq \mathcal{E}_\rho$, so it suffices to check $(\rho-\rho^{-1})/2 \geq C_0\rho$ for all $\rho \geq 3$. For $\rho = 3$, this follows by elementary calculations, and for $\rho > 3$, it follows by the fact that $\rho \mapsto (\rho-\rho^{-1})/(2\rho) = (1-\rho^{-2})/2$ is increasing for $\rho \geq 3$. □

**Theorem 15.3.7.** *Let $u$ be $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphic for some $\boldsymbol{b} \in \ell^p(\mathbb{N})$, $p \in (0,1)$, and $\varepsilon > 0$. Then there exists $J \in \mathbb{N}$ such that*

(i) *for each $\boldsymbol{\nu} \in \mathcal{F}$,*

$$c_{\boldsymbol{\nu}} := \int_{U_E} L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E) \frac{\partial^{\boldsymbol{\nu}_F}_{\boldsymbol{y}_F} u(\boldsymbol{y}_E, \boldsymbol{0})}{\boldsymbol{\nu}_F!} \mathrm{d}\mu_E(\boldsymbol{y}_E) \in \mathcal{X} \tag{15.20}$$

*is well-defined, and*

$$(\|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)} \|c_{\boldsymbol{\nu}}\|_{\mathcal{X}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^p(\mathcal{F}),$$

(ii) *we have*

$$u(\boldsymbol{y}) = \sum_{\boldsymbol{\nu}\in\mathcal{F}} c_{\boldsymbol{\nu}} L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E) \boldsymbol{y}_F^{\boldsymbol{\nu}_F} \in \mathcal{X}$$

*with absolute and uniform convergence for all $\boldsymbol{y} \in U$,*

(iii) *there exist constants $C_1, C_2 > 0$ and an increasing sequence $\boldsymbol{\delta} = (\delta_j)_{j\in\mathbb{N}} \subseteq (1,\infty)$ such that $(\delta_j^{-1})_{j\in\mathbb{N}} \in \ell^{p/(1-p)}(\mathbb{N})$, $\delta_j \leq C_1 j^{2/p}$ for all $j \in \mathbb{N}$ and*

$$(\boldsymbol{\delta}^{\boldsymbol{\nu}} \|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)} \|c_{\boldsymbol{\nu}}\|_{\mathcal{X}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^1(\mathcal{F}). \tag{15.21}$$

*Furthermore, with*

$$\Lambda_\tau := \{\boldsymbol{\nu} \in \mathcal{F} : \boldsymbol{\delta}^{-\boldsymbol{\nu}} \geq \tau\},$$

*we have for all $\tau \in (0,1)$ that $|\Lambda_\tau| > 0$ and*

$$\sup_{\boldsymbol{y}\in U} \left\| u(\boldsymbol{y}) - \sum_{\boldsymbol{\nu}\in\Lambda_\tau} c_{\boldsymbol{\nu}} L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E) \boldsymbol{y}_F^{\boldsymbol{\nu}_F} \right\|_{\mathcal{X}} \leq C_2 |\Lambda_\tau|^{-\frac{1}{p}+1}.$$

The proof is given in [61, Appendix A.1]. We next give more detail on the structure of the sets $(\Lambda_\tau)_{\tau\in(0,1)}$, which will be required in establishing the ensuing DNN expression rate bounds. To this end, let us introduce the quantities

$$m(\Lambda) := \sup_{\boldsymbol{\nu}\in\Lambda} |\boldsymbol{\nu}|_1 \quad \text{and} \quad d(\Lambda) := \sup_{\boldsymbol{\nu}\in\Lambda} |\operatorname{supp} \boldsymbol{\nu}|. \tag{15.22}$$

**Proposition 15.3.8.** *Let the assumptions of Theorem* 15.3.7 *be satisfied, and let $J \in \mathbb{N}$ and $(\Lambda_\tau)_\tau \in (0,1)$ be as in the statement of Theorem* 15.3.7. *Then*

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

(i) $\Lambda_\tau$ *is finite and downward closed for all* $\tau \in (0,1)$,
(ii) $m(\Lambda_\tau) = O(\log(|\Lambda_\tau|))$ *and* $d(\Lambda_\tau) = o(\log(|\Lambda_\tau|))$ *as* $\tau \to 0$,
(iii) $|\{\boldsymbol{\nu}_E : \boldsymbol{\nu} \in \Lambda_\tau\}| = O(\log(|\Lambda_\tau|)^J)$ *as* $\tau \to 0$,
(iv) *for all* $\tau \in (0,1)$, *if* $\mathbf{e}_j \in \Lambda_\tau$ *for some* $j \in \mathbb{N}$, *then* $\mathbf{e}_i \in \Lambda_\tau$ *for all* $i < j$.

*Proof.* To show (i), for downward closedness, let $\boldsymbol{\nu} \leq \boldsymbol{\mu}$ and $\boldsymbol{\mu} \in \Lambda_\tau$. Then $\tau \leq \boldsymbol{\rho}^{-\boldsymbol{\mu}} \leq \boldsymbol{\rho}^{-\boldsymbol{\nu}}$, and thus $\boldsymbol{\nu} \in \Lambda_\tau$. Item (ii) was shown in [81, Lemma 1.4.15] and [81, Example 1.4.23].

Item (iii) is a consequence of $m(\Lambda_\tau) = O(\log(|\Lambda_\tau|))$, which holds by (ii). Finally, (iv) is a direct consequence of the monotonicity of $(\delta_j)_{j\in\mathbb{N}}$, which holds by Theorem 15.3.7. □

**Remark 15.3.9.** We note that in the proof of Theorem 15.3.7, in particular, [61, Equation (A.9)], the sequence $\boldsymbol{\delta}$ is defined in terms of only $\boldsymbol{b}$, $p$, and $\varepsilon$.[2] The index sets $(\Lambda_\tau)_{\tau\in(0,1)}$ depend solely on $\boldsymbol{\delta}$ and $\tau$. Thus, in principle, $\varepsilon$ and the sequence $\boldsymbol{b}$ are sufficient to determine these index sets. For example, in the situation of Lemma 15.3.5, we have $b_j = \|\psi_j\|_{\mathcal{Z}}$, $j \in \mathbb{N}$, which is known (or can be estimated) for many function systems $\{\psi_j\}_{j\geq 1}$.

## 15.4 DNN surrogates of real-valued functions

We now turn to the statement and proofs of the main results of this work. We first recapitulate in Section 15.4.1 the DNNs, which we consider for approximation, and then present in Section 15.4.2 mathematical operations on DNNs. In Section 15.4.3, we recapitulate quantitative approximation rate bounds for polynomials by ReLU NNs from [45, 49, 62, 74], which we use subsequently to reapproximate $N$-term gPC approximations of $(\boldsymbol{b}, \varepsilon, \mathbb{R})$-holomorphic functions.

As in [74], we develop the DNN expression rate bounds (which are free from the curse of dimensionality of the parametric maps) in Sections 15.4.4 and 15.4.5 in an abstract setting for countably parametric scalar-valued maps with quantified control on the size of holomorphy domains.

### 15.4.1 Network architecture

We will use the same DNN architecture as in the previous works (e. g., [62]). In Sections 15.4.1–15.4.3, we now restate results from [62, Section 2].

We consider *deep neural networks (DNNs)* of feed-forward type. Such an NN $f$ can mathematically be described as a repeated composition of linear transformations with

**2** The sequence $\boldsymbol{\delta}$ depends on $\varepsilon$ through $\gamma_2 \in (1, \kappa)$ for $\kappa$ satisfying [61, Equation (A.1)].

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

a nonlinear activation function. More precisely: For an *activation function* $\sigma : \mathbb{R} \to \mathbb{R}$, a fixed *number of hidden layers* $L \in \mathbb{N}_0$, numbers $N_\ell \in \mathbb{N}$ of *computation nodes in layer* $\ell \in \{1, \dots, L+1\}$, $f : \mathbb{R}^{N_0} \to \mathbb{R}^{N_{L+1}}$ is realized by a feedforward neural network if for certain *weights* $w_{i,j}^\ell \in \mathbb{R}$ and *biases* $b_j^\ell \in \mathbb{R}$, we have, for all $\boldsymbol{x} = (x_i)_{i=1}^{N_0}$,

$$z_j^1 = \sigma\left(\sum_{i=1}^{N_0} w_{i,j}^1 x_i + b_j^1\right), \quad j \in \{1, \dots, N_1\}, \tag{15.23a}$$

$$z_j^{\ell+1} = \sigma\left(\sum_{i=1}^{N_\ell} w_{i,j}^{\ell+1} z_i^\ell + b_j^{\ell+1}\right), \quad \ell \in \{1, \dots, L-1\},\ j \in \{1, \dots, N_{\ell+1}\}, \tag{15.23b}$$

and, finally,

$$f(\boldsymbol{x}) = (z_j^{L+1})_{j=1}^{N_{L+1}} = \left(\sum_{i=1}^{N_L} w_{i,j}^{L+1} z_i^L + b_j^{L+1}\right)_{j=1}^{N_{L+1}}. \tag{15.23c}$$

In this case, $N_0$ is the dimension of the input, and $N_{L+1}$ is the dimension of the output. Furthermore, $z_j^\ell$ denotes the output of unit $j$ in layer $\ell$. The weight $w_{i,j}^\ell$ has the interpretation of connecting the $i$th unit in layer $\ell - 1$ with the $j$th unit in layer $\ell$. If $L = 0$, then (15.23c) holds with $z_i^0 := x_i$ for $i = 1, \dots, N_0$.

Except when explicitly stated, we will not distinguish between the network (which is defined through $\sigma$, $w_{i,j}^\ell$, and $b_j^\ell$) and the function $f : \mathbb{R}^{N_0} \to \mathbb{R}^{N_{L+1}}$ it realizes. We note in passing that this relation is typically not one-to-one, i. e., different NNs may realize the same function as their outputs. Let us also emphasize that we allow the weights $w_{i,j}^\ell$ and biases $b_j^\ell$ for $\ell \in \{1, \dots, L+1\}$, $i \in \{1, \dots, N_{\ell-1}\}$, and $j \in \{1, \dots, N_\ell\}$ to take any value in $\mathbb{R}$, i. e., we do not consider quantization as, e. g., in [10, 64].

As is customary in the theory of NNs, the number of hidden layers $L$ of an NN is referred to as *depth*,[3] and the total number of nonzero weights and biases as the *size* of the NN. Hence, for a DNN $f$ as in (15.23), we define

$$\operatorname{size}(f) := \big|\{(i,j,\ell) : w_{i,j}^\ell \neq 0\}\big| + \big|\{(j,\ell) : b_j^\ell \neq 0\}\big| \quad \text{and} \quad \operatorname{depth}(f) := L.$$

In addition, $\operatorname{size}_{\text{in}}(f) := |\{(i,j) : w_{i,j}^1 \neq 0\}| + |\{j : b_j^1 \neq 0\}|$ and $\operatorname{size}_{\text{out}}(f) := |\{(i,j) : w_{i,j}^{L+1} \neq 0\}| + |\{j : b_j^{L+1} \neq 0\}|$, which are the numbers of nonzero weights and biases in the input and output layers of $f$, respectively.

The proofs of our main results are constructive in the sense that we explicitly provide NN architectures and constructions of instances of DNNs with these architectures, which are sufficient (but possibly larger than necessary) for achieving the

**3** In other recent references (e. g., [60]), slightly different terminology for the number $L$ of layers in the DNN, differing from the convention in the present paper by a constant factor, is used. This difference will be inconsequential for all results that follow.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

claimed expression rates. We construct these NNs by assembling smaller networks using the operations of concatenation and parallelization, as well as so-called "identity networks", which realize the identity mapping. Below we recall the definitions.

## 15.4.2 Basic operations

Throughout, as activation function $\sigma$, we consider either the ReLU activation function

$$\sigma_1(x) := \max\{0, x\}, \quad x \in \mathbb{R}, \tag{15.24}$$

or, as suggested in [45, 52, 53], for $r \in \mathbb{N}$, $r \geq 2$, the RePU activation function

$$\sigma_r(x) := \max\{0, x\}^r = \sigma_1(x)^r, \quad x \in \mathbb{R}. \tag{15.25}$$

See [62, Remark 2.1] for a historical note on rectified power units. If an NN uses $\sigma_r$ as an activation function, then we refer to it as a $\sigma_r$-NN. ReLU NNs are referred to as $\sigma_1$-NNs. We assume throughout that all activations in a DNN are of equal type.

We now recall the parallelization and concatenation of networks, as well as networks realizing the identity. The constructions are mostly straightforward. For details and proofs, we refer to [26, 60, 62, 64].

### 15.4.2.1 Parallelization

Let $f, g$ be two NNs with the same depth $L \in \mathbb{N}_0$, input dimensions $n_f$, $n_g$ and output dimensions $m_f$, $m_g$ respectively. There exists an NN $(f,g)_{\mathrm{d}}$ such that

$$(f,g)_{\mathrm{d}} : \mathbb{R}^{n_f} \times \mathbb{R}^{n_g} \to \mathbb{R}^{m_f} \times \mathbb{R}^{m_g} : (\boldsymbol{x}, \tilde{\boldsymbol{x}}) \mapsto \big(f(\boldsymbol{x}), g(\tilde{\boldsymbol{x}})\big).$$

We have $\mathrm{depth}((f,g)_{\mathrm{d}}) = L$, $\mathrm{size}((f,g)_{\mathrm{d}}) = \mathrm{size}(f) + \mathrm{size}(g)$, $\mathrm{size}_{\mathrm{in}}((f,g)_{\mathrm{d}}) = \mathrm{size}_{\mathrm{in}}(f) + \mathrm{size}_{\mathrm{in}}(g)$, and $\mathrm{size}_{\mathrm{out}}((f,g)_{\mathrm{d}}) = \mathrm{size}_{\mathrm{out}}(f) + \mathrm{size}_{\mathrm{out}}(g)$; see [26, 64].

In case $n_f = n_g = n$, there exists an NN $(f,g)$ with the same depth and size as $(f,g)_d$, such that

$$(f,g) : \mathbb{R}^n \to \mathbb{R}^{m_f} \times \mathbb{R}^{m_g} : \boldsymbol{x} \mapsto \big(f(\boldsymbol{x}), g(\boldsymbol{x})\big).$$

### 15.4.2.2 Identity

By [64, Lemma 2.3], for all $n \in \mathbb{N}$ and $L \in \mathbb{N}_0$, there exists a $\sigma_1$-*identity network* $\mathrm{Id}_{\mathbb{R}^n}$ of depth $L$ such that $\mathrm{Id}_{\mathbb{R}^n}(\boldsymbol{x}) = \boldsymbol{x}$ for all $\boldsymbol{x} \in \mathbb{R}^n$. We have that

$$\mathrm{size}(\mathrm{Id}_{\mathbb{R}^n}) \leq 2n(L+1), \quad \mathrm{size}_{\mathrm{in}}(\mathrm{Id}_{\mathbb{R}^n}) \leq 2n, \quad \mathrm{size}_{\mathrm{out}}(\mathrm{Id}_{\mathbb{R}^n}) \leq 2n.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

Analogously, by [62, Proposition 2.3], for all $r, n \in \mathbb{N}$, $r \geq 2$, and $L \in \mathbb{N}_0$, there exists a *$\sigma_r$-identity network* $\mathrm{Id}_{\mathbb{R}^n}$ of depth $L$ such that $\mathrm{Id}_{\mathbb{R}^n}(\boldsymbol{x}) = \boldsymbol{x}$. We have that

$$\mathrm{size}(\mathrm{Id}_{\mathbb{R}^n}) \leq nL(4r^2 + 2r), \quad \mathrm{size}_{\mathrm{in}}(\mathrm{Id}_{\mathbb{R}^n}) \leq 4nr, \quad \mathrm{size}_{\mathrm{out}}(\mathrm{Id}_{\mathbb{R}^n}) \leq n(2r+1).$$

#### 15.4.2.3 Sparse concatenation

Let $f$ and $g$ be $\sigma_1$-NNs such that the output dimension of $g$ equals the input dimension of $f$. Let $n_g$ be the input dimension of $g$, and let $m_f$ be the output dimension of $f$. Then the *sparse concatenation of the NNs $f$ and $g$* realizes the function

$$f \circ g : \mathbb{R}^{n_g} \to \mathbb{R}^{m_f} : \boldsymbol{x} \mapsto f(g(\boldsymbol{x})). \tag{15.26}$$

In the following, by abuse of notation, "$\circ$" can either stand for the composition of functions or the sparse concatenation of networks. The meaning will be clear from the context. By [64, Remark 2.6], $\mathrm{depth}(f \circ g) = \mathrm{depth}(f) + 1 + \mathrm{depth}(g)$,

$$\mathrm{size}(f \circ g) \leq \mathrm{size}(f) + \mathrm{size}_{\mathrm{in}}(f) + \mathrm{size}_{\mathrm{out}}(g) + \mathrm{size}(g) \leq 2\,\mathrm{size}(f) + 2\,\mathrm{size}(g), \tag{15.27}$$

and

$$\mathrm{size}_{\mathrm{in}}(f \circ g) \leq \begin{cases} \mathrm{size}_{\mathrm{in}}(g), & \mathrm{depth}(g) \geq 1, \\ 2\,\mathrm{size}_{\mathrm{in}}(g), & \mathrm{depth}(g) = 0, \end{cases}$$

$$\mathrm{size}_{\mathrm{out}}(f \circ g) \leq \begin{cases} \mathrm{size}_{\mathrm{out}}(f), & \mathrm{depth}(f) \geq 1, \\ 2\,\mathrm{size}_{\mathrm{out}}(f), & \mathrm{depth}(f) = 0. \end{cases}$$

Similarly, for $r \geq 2$, there exists a sparse concatenation of $\sigma_r$-NNs (we denote the concatenation operator again by $\circ$) satisfying the following size and depth bounds from [62, Proposition 2.4]: Let $f$, $g$ be two $\sigma_r$-NNs such that the output dimension $k$ of $g$ equals the input dimension of $f$, and suppose that $\mathrm{size}_{\mathrm{in}}(f), \mathrm{size}_{\mathrm{out}}(g) \geq k$. Then $\mathrm{depth}(f \circ g) = \mathrm{depth}(f) + 1 + \mathrm{depth}(g)$,

$$\begin{aligned} \mathrm{size}(f \circ g) &\leq \mathrm{size}(f) + (2r-1)\,\mathrm{size}_{\mathrm{in}}(f) + (2r+1)k + (2r-1)\,\mathrm{size}_{\mathrm{out}}(g) + \mathrm{size}(g) \\ &\leq \mathrm{size}(f) + 2r\,\mathrm{size}_{\mathrm{in}}(f) + (4r-1)\,\mathrm{size}_{\mathrm{out}}(g) + \mathrm{size}(g) \\ &\leq (2r+1)\,\mathrm{size}(f) + 4r\,\mathrm{size}(g), \end{aligned} \tag{15.28}$$

and

$$\mathrm{size}_{\mathrm{in}}(f \circ g) \leq \begin{cases} \mathrm{size}_{\mathrm{in}}(g), & \mathrm{depth}(g) \geq 1, \\ 2r\,\mathrm{size}_{\mathrm{in}}(g) + 2rk \leq 4r\,\mathrm{size}_{\mathrm{in}}(g), & \mathrm{depth}(g) = 0, \end{cases}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$$\operatorname{size}_{\text{out}}(f \circ g) \leq \begin{cases} \operatorname{size}_{\text{out}}(f), & \operatorname{depth}(f) \geq 1, \\ 2r \operatorname{size}_{\text{out}}(f) + k \leq (2r+1) \operatorname{size}_{\text{out}}(f), & \operatorname{depth}(f) = 0. \end{cases}$$

Combining identity networks with the sparse concatenation, we can parallelize networks of different depth. The next lemma shows this for ReLU-NNs (a proof is given in [61, Appendix A.2]).

**Lemma 15.4.1.** *For all $k, n \in \mathbb{N}$ and $\sigma_1$-NNs $f_1, \dots, f_k$ with the same input dimension $n$ and output dimensions $m_1, \dots, m_k \in \mathbb{N}$, there exists a $\sigma_1$-NN $(f_1, \dots, f_k)_{\text{s}}$ called the* parallelization of $f_1, \dots, f_k$, with shared identity network. *It has input dimension $n$ and output dimension $m := \sum_{t=1}^k m_t$, realizes $\mathbb{R}^n \to \mathbb{R}^m : \boldsymbol{x} \mapsto (f_1(\boldsymbol{x}), \dots, f_k(\boldsymbol{x}))$, and has depth $L := \max_{t=1,\dots,k} \operatorname{depth}(f_t)$, and its size is bounded as follows:*

$$\begin{aligned} \operatorname{size}\big((f_1, \dots, f_k)_{\text{s}}\big) &\leq \sum_{t=1}^k \operatorname{size}(f_t) + \sum_{t=1}^k \operatorname{size}_{\text{in}}(f_t) + 2nL \leq 2 \sum_{t=1}^k \operatorname{size}(f_t) + 2nL, \\ \operatorname{size}_{\text{in}}\big((f_1, \dots, f_k)_{\text{s}}\big) &\leq \sum_{t=1}^k \operatorname{size}_{\text{in}}(f_t) + 2n, \\ \operatorname{size}_{\text{out}}\big((f_1, \dots, f_k)_{\text{s}}\big) &\leq \sum_{t=1}^k 2 \operatorname{size}_{\text{out}}(f_t). \end{aligned}$$

**Remark 15.4.2.** The term $2nL$ in the size bound corresponds to the nonzero weights (and biases) of the identity network used to construct the parallelization. We point out that this number is independent of the number $k$ of networks $(f_t)_{t=1}^k$, since our construction allows the $k$ networks to share one identity network.

## 15.4.3 Approximation of polynomials

As in other recent works (e. g., [22, 60, 62, 74]), the ensuing DNN expression rate analysis of possibly countably parametric posterior densities will rely on *DNN reapproximation of sparse generalized polynomial chaos approximations* of these densities. It has been observed in [49, 80] that ReLU DNNs can represent high-order polynomials on bounded intervals rather efficiently. We recapitulate several results of this type from [62, Section 2] and [74], which we will need in the following.

### 15.4.3.1 Approximate multiplication

Contrary to [80], the next result bounds the DNN expression error in $W^{1,\infty}([-M, M]^2)$ (instead of the $L^\infty([-M, M]^2)$-norm).

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Proposition 15.4.3** ([74, Proposition 3.1]). *For any $\delta \in (0,1)$ and $M \geq 1$, there exists a $\sigma_1$-NN $\tilde{\times}_{\delta,M} : [-M,M]^2 \to \mathbb{R}$ such that*

$$\begin{aligned} \sup_{|a|,|b|\leq M} \left|ab - \tilde{\times}_{\delta,M}(a,b)\right| &\leq \delta, \\ \operatorname*{ess\,sup}_{|a|,|b|\leq M} \max\left\{\left|b - \frac{\partial}{\partial a}\tilde{\times}_{\delta,M}(a,b)\right|, \left|a - \frac{\partial}{\partial b}\tilde{\times}_{\delta,M}(a,b)\right|\right\} &\leq \delta, \end{aligned} \tag{15.29}$$

*where $\frac{\partial}{\partial a}\tilde{\times}_{\delta,M}(a,b)$ and $\frac{\partial}{\partial b}\tilde{\times}_{\delta,M}(a,b)$ denote weak derivatives. There exists a constant $C > 0$, independent of $\delta \in (0,1)$ and $M \geq 1$, such that $\operatorname{size}_{\text{in}}(\tilde{\times}_{\delta,M}) \leq C$, $\operatorname{size}_{\text{out}}(\tilde{\times}_{\delta,M}) \leq C$,*

$$\operatorname{depth}(\tilde{\times}_{\delta,M}) \leq C(1 + \log_2(M/\delta)), \quad \operatorname{size}(\tilde{\times}_{\delta,M}) \leq C(1 + \log_2(M/\delta)).$$

*Moreover, for every $a \in [-M,M]$, there exists a finite set $\mathcal{N}_a \subseteq [-M,M]$ such that $b \mapsto \tilde{\times}_{\delta,M}(a,b)$ is strongly differentiable at all $b \in (-M,M)\backslash\mathcal{N}_a$.*

Proposition 15.4.3 implies the existence of networks approximating the multiplication of $n$ different numbers.

**Proposition 15.4.4** ([74, Proposition 3.3]). *For any $\delta \in (0,1)$, $n \in \mathbb{N}$, and $M \geq 1$, there exists a $\sigma_1$-NN $\tilde{\prod}_{\delta,M} : [-M,M]^n \to \mathbb{R}$ such that*

$$\sup_{(x_i)_{i=1}^n \in [-M,M]^n} \left|\prod_{j=1}^n x_j - \tilde{\prod}_{\delta,M}(x_1,\ldots,x_n)\right| \leq \delta. \tag{15.30}$$

*There exists a constant $C$, independent of $\delta \in (0,1)$, $n \in \mathbb{N}$, and $M \geq 1$, such that*

$$\begin{aligned} \operatorname{size}\left(\tilde{\prod}_{\delta,M}\right) &\leq C(1 + n\log(nM^n/\delta)), \\ \operatorname{depth}\left(\tilde{\prod}_{\delta,M}\right) &\leq C(1 + \log(n)\log(nM^n/\delta)). \end{aligned} \tag{15.31}$$

**Remark 15.4.5.** In [74], Propositions 15.4.3 and 15.4.4 are shown for $M = 1$. The result for $M > 1$ is obtained by a simple scaling argument. See [62, Proposition 2.6] for more detail.

### 15.4.3.2 ReLU DNN approximation of tensor product Legendre polynomials

Based on the ReLU DNN emulation of products in Proposition 15.4.3, we constructed ReLU DNN approximations of multivariate Legendre polynomials in [62]. For the statement, recall $m(\Lambda)$ in (15.22).

**Proposition 15.4.6** ([62, Proposition 2.13]). *For every finite $\Lambda \subset \mathbb{N}_0^d$ and every $\delta \in (0,1)$, there exists a $\sigma_1$-NN $f_{\Lambda,\delta} = (\tilde{L}_{\boldsymbol{\nu},\delta})_{\boldsymbol{\nu}\in\Lambda}$ with input dimension $d$ and output dimension $|\Lambda|$*

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*such that the outputs* $\{\tilde{L}_{\boldsymbol{\nu},\delta}\}_{\boldsymbol{\nu}\in\Lambda}$ *of* $f_{\Lambda,\delta}$ *satisfy, for every* $\boldsymbol{\nu} \in \Lambda$,

$$\|L_{\boldsymbol{\nu}} - \tilde{L}_{\boldsymbol{\nu},\delta}\|_{W^{1,\infty}([-1,1]^d)} \leq \delta, \quad \sup_{\boldsymbol{y}\in[-1,1]^d} |\tilde{L}_{\boldsymbol{\nu},\delta}((y_j)_{j\in\mathrm{supp}\,\boldsymbol{\nu}})| \leq \big(2m(\Lambda)+2\big)^d.$$

*Furthermore, there exists* $C > 0$ *such that for all* $d$, $\Lambda$, *and* $\delta$,

$$\begin{aligned}\mathrm{depth}(f_{\Lambda,\delta}) &\leq C(1+d\log d)\big(1+\log_2 m(\Lambda)\big)\big(m(\Lambda)+\log_2(1/\delta)\big),\\ \mathrm{size}(f_{\Lambda,\delta}) &\leq C\big[d^2 m(\Lambda)^2 + dm(\Lambda)\log_2(1/\delta) + d^2|\Lambda|\big(1+\log_2 m(\Lambda)+\log_2(1/\delta)\big)\big].\end{aligned}$$

#### 15.4.3.3 RePU DNN emulation of polynomials

The approximation of polynomials by neural networks can be significantly simplified if instead of the ReLU activation $\sigma_1$, we consider as an activation function the so-called *rectified power unit* (RePU) $\sigma_r(x) = \max\{0,x\}^r$ for $r \geq 2$. In contrast to $\sigma_1$-NNs, as shown in [45], for every $r \in \mathbb{N}$, $r \geq 2$, there exist RePU networks of depth 1 realizing the multiplication of two real numbers *without error*. This yields the following result, slightly improving [45, Theorem 9] in that the constant $C$ is independent of $d$. This is relevant, as in Section 15.4.5 the number of active parameters $d(\Lambda_\tau)$ increases with decreasing accuracy $\tau$.

**Proposition 15.4.7** ([62, Proposition 2.14]). *Fix* $d \in \mathbb{N}$ *and* $r \in \mathbb{N}$, $r \geq 2$. *Then there exists a constant* $C > 0$, *depending on* $r$ *but independent of* $d$, *such that for any finite downward closed* $\Lambda \subseteq \mathbb{N}_0^d$ *and any* $p \in \mathbb{P}_\Lambda$, *there is a* $\sigma_r$*-network* $\tilde{p} : \mathbb{R}^d \to \mathbb{R}$ *that realizes* $p$ *exactly and such that* $\mathrm{size}(\tilde{p}) \leq C|\Lambda|$ *and* $\mathrm{depth}(\tilde{p}) \leq C\log_2(|\Lambda|)$.

**Remark 15.4.8.** Similar results hold for other widely used activation functions $\psi$. As discussed in [62, Remark 2.15], if the product of two numbers can be approximated by $\psi$-NNs up to *arbitrary accuracy* and with NN size and depth *independent of the accuracy*, then polynomials can be approximated with size and depth bounded as $\mathrm{size}(\tilde{p}) \leq C|\Lambda|$ and $\mathrm{depth}(\tilde{p}) \leq C\log_2(|\Lambda|)$ for $C$ independent of the arbitrarily small accuracy.

Activation functions for which this holds include (i) $\psi \in C^2$ for which there exists $x \in \mathbb{R}$ such that $\psi''(x) \neq 0$, (ii) $\psi$ that are continuous and *sigmoidal of order* $k \geq 2$ (see also [62, Remark 2.1]), and (iii) NNs with rational activations. We refer to [62, Remark 2.15] for a more detailed discussion.

### 15.4.4 ReLU DNN approximation of $(\boldsymbol{b},\boldsymbol{\varepsilon},\mathbb{R})$-holomorphic maps

We now present a result about the expressive power for $(\boldsymbol{b},\varepsilon,\mathbb{R})$-holomorphic functions in the sense of Remark 15.3.1. Theorem 15.4.9 generalizes [74, Theorem 3.9], as it

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

shows that less regular functions[4] can be emulated with the same convergence rate (see Remark 15.3.4). In particular, we obtain that up to logarithmic terms, ReLU DNNs are capable of approximating $(\boldsymbol{b}, \varepsilon, \mathbb{R})$-holomorphic maps at rates equivalent to those achieved by best $n$-term gPC approximations. Here *"rate" is understood in terms of the NN size*, i. e., in terms of the total number of nonzero weights in the DNN.

In the following, for $\Lambda_\tau \subset \mathcal{F}$ as in Theorem 15.3.7, we define its support

$$S_{\Lambda_\tau} := \bigcup_{\boldsymbol{\nu}\in\Lambda_\tau} \operatorname{supp} \boldsymbol{\nu} \subset \mathbb{N}. \tag{15.32}$$

**Theorem 15.4.9.** *Let $u : U \to \mathbb{R}$ be $(\boldsymbol{b}, \varepsilon, \mathbb{R})$-holomorphic for some $\boldsymbol{b} \in \ell^p(\mathbb{N})$, $p \in (0,1)$, and $\varepsilon > 0$. For $\tau \in (0,1)$, let $\Lambda_\tau \subset \mathcal{F}$ be as in Theorem* 15.3.7.

*Then there exists $C > 0$ depending on $\boldsymbol{b}$, $\varepsilon$, and $u$, such that for all $\tau \in (0,1)$, there exists a $\sigma_1$-NN $\tilde{u}_\tau$ with input variables $(y_j)_{j\in S_{\Lambda_\tau}}$ such that*

$$\begin{aligned} \operatorname{size}(\tilde{u}_\tau) &\le C\big(1 + |\Lambda_\tau| \cdot \log|\Lambda_\tau| \cdot \log\log|\Lambda_\tau|\big), \\ \operatorname{depth}(\tilde{u}_\tau) &\le C\big(1 + \log|\Lambda_\tau| \cdot \log\log|\Lambda_\tau|\big). \end{aligned}$$

*Furthermore, $\tilde{u}_\tau$ satisfies the uniform error bound*

$$\sup_{\boldsymbol{y}\in U} \big|u(\boldsymbol{y}) - \tilde{u}_\tau\big((y_j)_{j\in S_{\Lambda_\tau}}\big)\big| \le C|\Lambda_\tau|^{-1/p+1}. \tag{15.33}$$

*In the case $|\Lambda_\tau| = 1$, the statement holds with* $\log\log|\Lambda_\tau|$ *replaced by* $0$.

The proof is given in [61, Appendix A.3].

**Remark 15.4.10.** Let $K \in \mathbb{N}$, and let $v : U \to \mathbb{R}^K$ be $(\boldsymbol{b}, \varepsilon, \mathbb{R}^K)$-holomorphic. Then Theorem 15.4.9 can be applied to each component of $v$. This at most increases the bound on the network size by a factor $K$, but it does not affect the depth and convergence rate. In fact, only the dimension of the output layer has to be increased, since the hidden layers of the DNN can be the same as for $K = 1$. This corresponds to reusing the same polynomial basis for the approximation of all components of $v$.

## 15.4.5 RePU DNN approximation of $(\boldsymbol{b}, \boldsymbol{\varepsilon}, \mathbb{R})$-holomorphic maps

We next provide an analogue of Theorem 15.4.9 (which used $\sigma_1$-NNs) for $\sigma_r$-NNs, $r \ge 2$. The smaller multiplication networks of Proposition 15.4.7 allow us to prove the same approximation error for slightly smaller networks in this case.

**4** Theorem 15.4.9 only assumes quantified holomorphy in *polyellipses* in a suitable finite number of the parameters $y_j$, whereas [74, Theorem 3.9] required holomorphy in poly*discs*. The presently obtained expression rates are identical to those in [74, Theorem 3.9] but are shown to hold for maps with smaller domains of holomorphy.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Theorem 15.4.11.** *Let $u : U \to \mathbb{R}$ be $(\boldsymbol{b}, \varepsilon, \mathbb{R})$-holomorphic for some $\boldsymbol{b} \in \ell^p(\mathbb{N})$, $p \in (0,1)$, and $\varepsilon > 0$. For $\tau \in (0,1)$, let $\Lambda_\tau \subset \mathcal{F}$ be as in Theorem* 15.3.7. *Let $r \in \mathbb{N}$, $r \geq 2$.*

*Then there exists $C > 0$ depending on $\boldsymbol{b}$, $\varepsilon$, $u$, and $r$ such that for all $\tau \in (0,1)$, there exists a $\sigma_r$-NN $\tilde{u}_\tau$ with input variables $(y_j)_{j \in S_{\Lambda_\tau}}$ such that*

$$\operatorname{size}(\tilde{u}_\tau) \leq C|\Lambda_\tau|, \quad \operatorname{depth}(\tilde{u}_\tau) \leq C \log|\Lambda_\tau|,$$

*and $\tilde{u}_\tau$ satisfies the uniform error bound*

$$\sup_{\boldsymbol{y} \in U} \left| u(\boldsymbol{y}) - \tilde{u}_\tau((y_j)_{j \in S_{\Lambda_\tau}}) \right| \leq C|\Lambda_\tau|^{-1/p+1}. \tag{15.34}$$

*Proof.* By Proposition 15.4.7 the $|S_{\Lambda_\tau}|$-variate polynomial $\sum_{\boldsymbol{\nu} \in \Lambda_\tau} c_{\boldsymbol{\nu}} L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E) \boldsymbol{y}_F^{\boldsymbol{\nu}_F} \in \mathbb{P}_{\Lambda_\tau}$ from Theorem 15.3.7 and Corollary 15.3.8 can be emulated exactly by a $\sigma_r$-NN satisfying

$$\operatorname{size}(\tilde{u}_\tau) \leq C|\Lambda_\tau|, \quad \operatorname{depth}(\tilde{u}_\tau) \leq C \log(|\Lambda_\tau|)$$

for $C$ independent of $|S_{\Lambda_\tau}|$. The error bound (15.34) holds by Theorem 15.3.7(iii). □

Remarks 15.4.8 and 15.4.10 also apply here.

# 15.5 DNN surrogates of $\mathcal{X}$-valued functions

In this section, we address the DNN emulation of countably parametric holomorphic maps taking values in function spaces as typically arise in PDE UQ. In Section 15.5.1, we show DNN expression rate bounds for parametric PDE solution families, assuming the existence of suitable NN approximations of functions in the solution space of the PDE.

In Section 15.5.2.1, we review results on the *exact DNN emulation of Courant-type finite element spaces* on regular simplicial triangulations. In Sections 15.5.2.2 and 15.5.2.3, we discuss Theorem 15.5.2 for the diffusion equation from Section 15.2.4.1 and the eigenvalue problem from Section 15.2.4.2.

## 15.5.1 ReLU DNN expression of $(\boldsymbol{b}, \boldsymbol{\varepsilon}, \mathcal{X})$-holomorphic maps

So far, we considered the DNN expression of real-valued maps $u : U \to \mathbb{R}$. In applications to PDEs, often also the expression of maps $u : U \to \mathcal{X}$ is of interest. Here the real Banach space $\mathcal{X}$ is a function space over a domain $\mathrm{D} \subset \mathbb{R}^d$ for $d \in \mathbb{N}$ and is interpreted as the solution space of the parametric forward model (15.1).

As it was shown, for example, in [3, 25, 82], for the gPC coefficients $u_{\boldsymbol{\nu}}$, a $\boldsymbol{\nu}$-dependent degree of resolution in $\mathcal{X}$ of $u_{\boldsymbol{\nu}}$ is in general advantageous. We approach DNN

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

expression of the parametric solution map through DNN emulation of multilevel gPC-FE approximations. To state these, we need a *regularity space* $\mathcal{X}^s \subset \mathcal{X}$ of functions with additional regularity. We first present the result in an abstract setting and subsequently detail it for an example in Sections 15.5.2.2 and 15.5.2.3.

For the DNN emulation of polynomials in the variables $\boldsymbol{y} \in U$, we use [61, Lemma A.1], based on the networks constructed in the proof of Theorem 15.4.9. For the gPC coefficients, which we assume to be in $\mathcal{X}^s$, we allow sequences of NN approximations satisfying a mild bound on their $L^\infty$-norm, as made precise in Assumption 15.5.1. This is needed to use the product networks from Proposition 15.4.3 to multiply NNs approximating the polynomials in $\boldsymbol{y}$ with NN approximations of gPC-coefficients.

**Assumption 15.5.1.** Assume that there exist $\gamma > 0$, $\theta \geq 0$, and $C > 0$ such that for all $v \in \mathcal{X}^s$ and $m \in \mathbb{N}$, there exists an NN $\Phi_m^v$ that satisfies

$$\operatorname{depth}(\Phi_m^v) \leq C(1 + \log m), \quad \operatorname{size}(\Phi_m^v) \leq Cm,$$

and

$$\|v - \Phi_m^v\|_{\mathcal{X}} \leq C\|v\|_{\mathcal{X}^s} m^{-\gamma}, \quad \|\Phi_m^v\|_{\mathcal{X}} \leq C\|v\|_{\mathcal{X}}, \quad \|\Phi_m^v\|_{L^\infty(\mathrm{D})} \leq C\|v\|_{\mathcal{X}^s} m^{\theta}.$$

Let us consider an example. For a bounded polytope $\mathrm{D} \subset \mathbb{R}^d$, functions in the Kondratiev space $\mathcal{X}^s = \mathcal{K}^2_{1+\zeta}(\mathrm{D})$ with $\zeta \in (0,1)$ (for a definition of $\mathcal{K}^2_{1+\zeta}(\mathrm{D})$, see Equation (15.39)) can be approximated by continuous piecewise affine functions on regular triangulations of D with convergence rate $\gamma = \frac{1}{d}$ (e. g., [2, 6, 46] for $d = 2$ and [58] for $d > 2$). Continuous piecewise affine functions on regular simplicial partitions can be exactly emulated by ReLU networks; see Section 15.5.2.1. These NNs approximate functions in $\mathcal{X}^s = \mathcal{K}^2_{1+\zeta}(\mathrm{D})$ with (optimal) rate $\gamma = 1/d$. By the continuous embedding $\mathcal{X}^s \hookrightarrow L^\infty(\mathrm{D})$ ([51], [19, Theorem 27]) the last inequality in Assumption 15.5.1 is satisfied with $\theta = 0$. Here the domain D may, but need not, be the physical domain of interest. The theorem below also applies to boundary integral equations, in which case D is the boundary of the physical domain. Holomorphic dependence of boundary integral operators on the shape of the domain ("shape-holomorphy") is shown in [32].

We obtain the following result, which generalizes [74, Theorem 4.8]. To state the theorem, we recall the notation $S_{\Lambda_\tau} = \bigcup_{\boldsymbol{\nu} \in \Lambda_\tau} \operatorname{supp} \boldsymbol{\nu} \subset \mathbb{N}$ introduced in (15.32).

**Theorem 15.5.2.** *Let $d \in \mathbb{N}$, and let $\mathcal{X} = W^{1,q}(\mathrm{D})$, $q \in [1,\infty]$,[5] and $\mathcal{X}^s \subset \mathcal{X}$ be Banach spaces of functions $v : \mathrm{D} \to \mathbb{R}$ for some bounded domain $\mathrm{D} \subset \mathbb{R}^d$. Assume that Assumption 15.5.1 holds for some $\gamma > 0$ and $\theta \geq 0$. Let $u : U \to \mathcal{X}^s \subset \mathcal{X}$ be a $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphic map, in the sense of Remark 15.3.1, for some $\boldsymbol{b} \in \ell^p(\mathbb{N})$, $p \in (0,1)$, and*

**5** Although $q = 2$ in all examples we consider, the theorem is stated slightly more generally for $q \in [1,\infty]$. In fact, the result also holds for weighted $W^{1,q}$-spaces.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*$\varepsilon > 0$. Let $(c_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \subset \mathcal{X}$ and $(\Lambda_\tau)_{\tau\in(0,1)}$ be as in Theorem* 15.3.7. *Assume that $(c_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \subset \mathcal{X}^s$ and $(\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}^s}\|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^{p^s}$ for some $0 < p < p^s < 1$.*

*Then there exists a constant $C > 0$ depending on $d, \gamma, \theta, \boldsymbol{b}$ (thus also on $p$), $\varepsilon, p^s$, and $u$ such that for all $\tau \in (0,1)$, there exists a ReLU NN $\tilde{u}_\tau$ with input variables $(x_1, \dots, x_d) = \boldsymbol{x} \in \mathrm{D}$ and $(y_j)_{j\in S_{\Lambda_\tau}}$ for $\boldsymbol{y} \in U$ and output dimension 1 such that for some $\mathcal{N}_\tau \in \mathbb{N}$ satisfying $\mathcal{N}_\tau \geq |\Lambda_\tau|$,*

$$\mathrm{size}(\tilde{u}_\tau) \leq C(1 + \mathcal{N}_\tau \cdot \log \mathcal{N}_\tau \cdot \log\log \mathcal{N}_\tau), \quad \mathrm{depth}(\tilde{u}_\tau) \leq C(1 + \log \mathcal{N}_\tau \cdot \log\log \mathcal{N}_\tau)$$

*and such that $\tilde{u}_\tau$ satisfies the uniform error bound*

$$\sup_{\boldsymbol{y}\in U} \|u(\boldsymbol{y}) - \tilde{u}_\tau(\cdot, (y_j)_{j\in S_{\Lambda_\tau}})\|_{\mathcal{X}} \leq C\mathcal{N}_\tau^{-r}, \quad r := \gamma \min\left\{1, \frac{1/p - 1}{\gamma + 1/p - 1/p^s}\right\}. \tag{15.35}$$

The proof is given in [61, Appendix A.4]. Theorem 15.5.2 shows that for all $r^* < r$, there exists $C > 0$ (additionally depending on $r^*$) such that

$$\sup_{\boldsymbol{y}\in U} \|u(\boldsymbol{y}) - \tilde{u}_\tau(\cdot, (y_j)_{j\in S_{\Lambda_\tau}})\|_{\mathcal{X}} \leq C(\mathrm{size}(\tilde{u}_\tau))^{-r^*}.$$

The limit $r$ on the convergence rate in (15.35) is bounded from above by the gPC best $n$-term rate $1/p - 1$ for the truncation error of the gPC expansion and by the convergence rate $\gamma$ of ReLU DNN approximations of functions in $\mathcal{X}^s$ from Assumption 15.5.1.

## 15.5.2 ReLU DNN expression of Courant finite elements

We now recall that any continuous piecewise affine function on a locally convex regular triangulation is representable by a ReLU network; see, e. g., [31, 74]. This is used in Section 15.5.2.2 to show an expression for $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphic functions, where $\mathcal{X}$ is a Sobolev space over a bounded domain.

### 15.5.2.1 Continuous piecewise affine functions

In space dimension $d = 1$, any continuous piecewise linear function on a partition $a = t_0 < t_1 < \cdots < t_N = b$ of a finite interval $[a, b]$ into $N$ subintervals can be expressed without error by a $\sigma_1$-NN with depth 1 and size $\mathcal{O}(N)$; see, e. g., [74, Lemma 4.5].

A similar result holds for $d \geq 2$. Consider a bounded polytope $G \subset \mathbb{R}^d$ with Lipschitz boundary $\partial G$ being (the closure of) a finite union of plane $(d-1)$-faces. Let $\mathcal{T}$ be a regular simplicial triangulation of $G$, i. e., the intersection of any two distinct closed simplices $\overline{T}, \overline{T}' \in \mathcal{T}$ is either empty or an entire $k$-simplex for some $0 \leq k < d$.[6] For

**6** In other words, $\mathcal{T}$ is a cellular complex.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

the ReLU NN emulation of gPC-coefficients, we will use that also in space dimension $d \geq 2$, continuous piecewise linear functions on a regular simplicial mesh $\mathcal{T}$ can efficiently be emulated exactly by ReLU DNNs. For locally convex partitions, this was shown in [31], as we next recall in Proposition 15.5.3. The term *locally convex* refers to meshes $\mathcal{T}$ for which each patch consisting of all elements attached to a fixed node of $\mathcal{T}$ is a convex set. See [31] for more detail.

Set

$$S^1(G,\mathcal{T}) := \{v \in C^0(G) : v|_T \in \mathbb{P}_1 \; \forall T \in \mathcal{T}\}.$$

We denote by $\mathcal{N}(\mathcal{T})$ the set of nodes of the mesh $\mathcal{T}$ and by $k_{\mathcal{T}} := \max_{p\in\mathcal{N}} |\{T \in \mathcal{T} : p \in \overline{T}\}|$, the maximum number of elements sharing a node.

**Proposition 15.5.3** ([31, Theorem 3.1]). *Let $\mathcal{T}$ be a regular simplicial locally convex triangulation of a bounded polytope $G$. Then every $v \in S^1(G,\mathcal{T})$ can be implemented exactly by a $\sigma_1$-NN of depth $1 + \log_2\lceil k_{\mathcal{T}}\rceil$ and size of the order $\mathcal{O}(|\mathcal{T}|k_{\mathcal{T}})$.*

Estimates on the network size for continuous piecewise linear functions on general regular simplicial partitions $\mathcal{T}$ are stated in [31, Theorem 5.2] based on [78] but are much larger than those in [31, Theorem 3.1].

#### 15.5.2.2 Parametric diffusion problem

The standard example of a $(\boldsymbol{b},\varepsilon,\mathcal{X})$-holomorphic parametric solution family is based on Section 15.2.4.1, i. e., the solution to an affine-parametric diffusion problem; see, e. g., [13, 82]. In the setting of Section 15.2.4.1, we verify the assumptions of Theorem 15.5.2.

Let $\mathrm{D} \subset \mathbb{R}^2$ be a bounded polygonal Lipschitz domain (for details, see [81, Remark 4.2.1]). We consider a linear elliptic diffusion equation with uncertain diffusion coefficient and with homogeneous Dirichlet boundary conditions. With $\mathcal{X} = \mathcal{Y} := H^1_0(\mathrm{D};\mathbb{C})$ and $X := L^\infty(\mathrm{D};\mathbb{C})$, for a fixed right-hand side $f \in \mathcal{Y}' = \mathcal{X}'$, the weak formulation reads as follows: given $a \in X$, find $u(a) \in \mathcal{X}$ such that

$$\int_{\mathrm{D}} \nabla u(a)^\top a \nabla v \mathrm{d}\boldsymbol{x} = \langle f, v\rangle \quad \forall v \in \mathcal{Y}. \tag{15.36}$$

Then the map $G : a \mapsto u(a) \in \mathcal{X}$ is locally well-defined and holomorphic around every $a \in X$ for which $\operatorname{ess\,inf}_{\boldsymbol{x}\in\mathrm{D}} \Re(a(\boldsymbol{x})) > 0$; see, e. g., [81, Example 1.2.38 and Equations (4.3.12)–(4.3.13)].

We consider affine-parametric diffusion coefficients $a = a(\boldsymbol{y})$, where $\boldsymbol{y} = (y_j)_{j\in\mathbb{N}}$ is a sequence of real-valued parameters ranging in $U = [-1,1]^{\mathbb{N}}$. For a nominal input

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$a_0 \in X$ and for a sequence of fluctuations $(\psi_j)_{j\in\mathbb{N}} \subseteq X$, define

$$a(\boldsymbol{y}) = a_0 + \sum_{j\in\mathbb{N}} y_j\psi_j. \tag{15.37}$$

Such expansions arise, for example, from the Fourier, Karhunen–Loève, spline, or wavelet series representations of $a$.

If $\operatorname{ess\,inf}_{\boldsymbol{x}\in\mathrm{D}} \Re(a_0(\boldsymbol{x})) = \gamma > 0$, then

$$\sum_{j\in\mathbb{N}} \|\psi_j\|_X < \gamma \tag{15.38}$$

ensures $\operatorname{ess\,inf}_{\boldsymbol{x}\in\mathrm{D}} \Re(a(\boldsymbol{y})(\boldsymbol{x})) > 0$ for all $\boldsymbol{y} \in U$. This in turn implies that (15.36) admits a unique solution for all diffusion coefficients $a(\boldsymbol{y})$, $\boldsymbol{y} \in U$. Thus Lemma 15.3.5 yields $\boldsymbol{y} \mapsto u(\boldsymbol{y}) = G(a_0 + \sum_{j\in\mathbb{N}} y_j\psi_j)$ to be $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphic for some $\varepsilon > 0$ and with $b_j := \|\psi_j\|_X$, $j \in \mathbb{N}$.

Next, we consider a smoothness space $\mathcal{X}^s$ and recall $(\boldsymbol{b}^s, \varepsilon^s, \mathcal{X}^s)$-holomorphy of $u : U \to \mathcal{X}^s : \boldsymbol{y} \mapsto u(\boldsymbol{y})$. First, we recall the definition of Kondratiev spaces: Let $k \in \mathbb{N}_0$ and $\zeta \in \mathbb{R}$, and let $r_{\mathrm{D}} : \mathrm{D} \to \mathbb{R}_{>0}$ be a smooth function which near vertices of D equals the distance to the closest vertex. Then

$$\mathcal{K}^k_\zeta(\mathrm{D}) := \{u : \mathrm{D} \to \mathbb{C} : r_{\mathrm{D}}^{|\boldsymbol{\xi}|-\zeta}\partial_x^{\boldsymbol{\xi}} u \in L^2(\mathrm{D}), \boldsymbol{\xi} \in \mathbb{N}_0^2, |\boldsymbol{\xi}| \le k\}. \tag{15.39}$$

To obtain the approximation rate $\gamma = \frac{1}{2}$ in Proposition 15.5.3, we consider $\mathcal{X}^s := \mathcal{K}^2_{\zeta+1}(\mathrm{D})$ for some $\zeta \in (0,1)$. By [5, Theorem 1.1] and [81, Example 1.2.38] there exists $\zeta \in (0,1)$ such that when $f \in \mathcal{K}^0_{\zeta-1}(\mathrm{D})$, $a \in W^{1,\infty}(\mathrm{D}) =: X^s$, and $\operatorname{ess\,inf}_{\boldsymbol{x}\in\mathrm{D}} \Re(a(\boldsymbol{x})) > 0$, the map $G : a \mapsto u(a) \in \mathcal{X}^s$ is locally well-defined and holomorphic around every such $a$. We remark that the space from which we chose $f$ satisfies $L^2(\mathrm{D}) \subset \mathcal{K}^0_{\zeta-1}(\mathrm{D}) \subset H^{-1}(\mathrm{D}) = \mathcal{Y}'$.

If in addition to previously made assumptions, $\{\psi_j\}_{j\in\mathbb{N}}$ satisfies

$$\sum_{j\in\mathbb{N}} \|\psi_j\|_{X^s} < \infty,$$

then Lemma 15.3.5 yields $\boldsymbol{y} \mapsto u(\boldsymbol{y}) = G(a_0 + \sum_{j\in\mathbb{N}} y_j\psi_j)$ to be $(\boldsymbol{b}^s, \varepsilon^s, \mathcal{X}^s)$-holomorphic for some $\varepsilon^s > 0$ and with $b_j^s := \|\psi_j\|_{X^s}$, $j \in \mathbb{N}$. For a more detailed discussion of this example and more general advection–diffusion–reaction equations, see [81, Section 4.3].

Thus, for the map $U \to \mathcal{X}^s \subset \mathcal{X} : \boldsymbol{y} \mapsto u(\boldsymbol{y})$ to be $(\boldsymbol{b}, \varepsilon, \mathcal{X})$- and $(\boldsymbol{b}^s, \varepsilon^s, \mathcal{X}^s)$-holomorphic for $\boldsymbol{b} \in \ell^p(\mathbb{N})$ and $\boldsymbol{b}^s \in \ell^{p^s}(\mathbb{N})$ for some $0 < p < p^s < 1$, we additionally need to assume that $(\|\psi_j\|_X)_{j\in\mathbb{N}} \in \ell^p(\mathbb{N})$ and $(\|\psi_j\|_{X^s})_{j\in\mathbb{N}} \in \ell^{p^s}(\mathbb{N})$. The $(\boldsymbol{b}^s, \varepsilon^s, \mathcal{X}^s)$-holomorphy and Theorem 15.3.7 give $(\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}^s}\|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^{p^s}$.

In summary, the assumptions on $u$ in Theorem 15.5.2 hold when $f \in L^2(\mathrm{D})$ and $a_0, \{\psi_j\}_{j\in\mathbb{N}} \subset W^{1,\infty}(\mathrm{D})$ satisfy $\operatorname{ess\,inf}_{\boldsymbol{x}\in\mathrm{D}} \Re(a_0(\boldsymbol{x})) > 0$, Equation (15.38), $(\|\psi_j\|_X)_{j\in\mathbb{N}} \in$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

$\ell^p(\mathbb{N})$, and $(\|\psi_j\|_{X^s})_{j\in\mathbb{N}} \in \ell^{p^s}(\mathbb{N})$. Then $u : U \to \mathcal{X}^s = \mathcal{K}^2_{\zeta+1}(\mathrm{D})$ for some $\zeta \in (0,1)$. As mentioned below Assumption 15.5.1, the NN approximations in Section 15.5.2.1 satisfy Assumption 15.5.1 with $\theta = 0$ and approximation rate $\gamma = \frac{1}{2}$.

#### 15.5.2.3 Parametric eigenvalue problem

We verify the assumptions of Theorem 15.5.2 for the parametric eigenvalue problem (15.12). To this end, we choose $\mathcal{X} := \mathbb{C} \times H^1_0(\mathrm{D};\mathbb{C})$, $X := L^\infty(\mathrm{D};\mathbb{C})$.

Then the parametric first eigenpair $\{(\lambda_1(\boldsymbol{y}), w_1(\boldsymbol{y})) : \boldsymbol{y} \in U\} \subset \mathcal{X}$ admits a unique holomorphic continuation $\{(\lambda_1(\boldsymbol{z}), w_1(\boldsymbol{z})) : \boldsymbol{z} \in V\} \subset \mathcal{X}$ to an open neighborhood $V$ of $U$ in $\mathbb{C}^{\mathbb{N}}$. The proof follows from the uniformity of the spectral gap of the parametric first and second eigenvalues, i. e., from $\lambda_2(\boldsymbol{y}) - \lambda_1(\boldsymbol{y}) > c_0$ for all $\boldsymbol{y} \in U$ and some $c_0 > 0$, which is shown in [28, Proposition 2.4]. Also, see [1, Theorem 4] for a proof of analytic dependence on each $y_j$. Upon defining the parametric "right-hand side" $f(\boldsymbol{y}) := \lambda_1(\boldsymbol{y})w_1(\boldsymbol{y}) \in H^1_0(\mathrm{D};\mathbb{R}) \subset L^2(\mathrm{D})$ for $\boldsymbol{y} \in U$, it follows that the map $u := (\lambda_1, w_1) \in \mathcal{X}$ satisfies $u : U \to \mathcal{X}^s = \mathbb{C} \times \mathcal{K}^2_{\zeta+1}(\mathrm{D})$ for some $\zeta \in (0,1)$. It is, in addition, $(\boldsymbol{b}, \varepsilon, \mathcal{X})$- and $(\boldsymbol{b}^s, \varepsilon^s, \mathcal{X}^s)$-holomorphic for $\boldsymbol{b} \in \ell^p(\mathbb{N})$ and $\boldsymbol{b}^s \in \ell^{p^s}(\mathbb{N})$ for some $0 < p < p^s < 1$, $\varepsilon, \varepsilon^s > 0$, provided that $(\|\psi_j\|_X)_{j\in\mathbb{N}} \in \ell^p(\mathbb{N})$ and $(\|\psi_j\|_{X^s})_{j\in\mathbb{N}} \in \ell^{p^s}(\mathbb{N})$. As before, $X^s = W^{1,\infty}(\mathrm{D})$. This $(\boldsymbol{b}, \varepsilon, \mathcal{X})$- and $(\boldsymbol{b}^s, \varepsilon^s, \mathcal{X}^s)$-holomorphy was proved in [1, Theorem 4 and Corollary 2] (where for simplicity the corollary was stated for the particular case of convex D).

## 15.6 Application to Bayesian inference

### 15.6.1 ReLU DNN approximations for inverse UQ

In this section, we discuss how the results in Section 15.4.4 apply to Bayesian inverse problems from Sections 15.2.1 and 15.2.2.1.

In practice, it is more convenient to work with measures on $U$, instead of their pushforwards under the map $\boldsymbol{y} \mapsto a(\boldsymbol{y}) := a_0 + \sum_{j\in\mathbb{N}} y_j\psi_j \in X$ on the Banach space $X$. For this reason, throughout this section, we adopt the equivalent viewpoint of interpreting $\boldsymbol{y} \in U$ (instead of $a(\boldsymbol{y})$) as the unknown, $\mu_U$ as the prior, and $a^{-1}\sharp\mu^\delta$ as the posterior measure on $U$ (which is the measure of the unknown $\boldsymbol{y} \in U$ conditioned on the data $\delta$). Here we assume that $a : \boldsymbol{y} \mapsto a(\boldsymbol{y})$ is invertible and that $a^{-1}$ is measurable, and denote by $a^{-1}\sharp\mu^\delta$ the pushforward measure of $\mu^\delta$ under $a^{-1}$ (which is a measure on $U$).[7]

7 Alternative to looking for the unknown $a(\boldsymbol{y})$ in the Banach space $X$, we could interpret $\boldsymbol{y} \in U$ to be the unknown. In this case the posterior measure is defined on $U$ (instead of $X$), and the assumption of invertibility of $a$, which is used to push forward $\mu^\delta$ to a measure on $U$, would not be necessary.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

**Corollary 15.6.1.** *Let $u$ be $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphic, let $\boldsymbol{b} \in \ell^p$, $p \in (0,1)$, and assume that the observation noise covariance $\Gamma \in \mathbb{R}^{K\times K}$ is symmetric and positive definite. Let the observation operator $\mathcal{O} : \mathcal{X} \to \mathbb{R}^K$ be deterministic, bounded, and linear, let $\mu_U$ be the uniform measure on $U = [-1,1]^{\mathbb{N}}$, and let for a given data sample $\delta \in \mathbb{R}^K$,*

$$\frac{\mathrm{d}a^{-1}\sharp\mu^\delta}{\mathrm{d}\mu_U}(\boldsymbol{y}) = \frac{1}{Z(\delta)}\exp\Big(-\frac{1}{2}\big\|\delta - \mathcal{O}(u(\boldsymbol{y}))\big\|_\Gamma^2\Big) \quad \text{for all } \boldsymbol{y} \in U,$$
$$Z(\delta) = \int_U \exp\Big(-\frac{1}{2}\big\|\delta - \mathcal{O}(u(\boldsymbol{y}))\big\|_\Gamma^2\Big)\mathrm{d}\mu_U(\boldsymbol{y}).$$

*Then also $\frac{\mathrm{d}a^{-1}\sharp\mu^\delta}{\mathrm{d}\mu_U}(\boldsymbol{y})$ is $(\boldsymbol{b}, \varepsilon, \mathbb{R})$-holomorphic.*

*By Theorem 15.4.9 it can thus be uniformly approximated by ReLU NNs with convergence rate (in terms of the size of the network) arbitrarily close to $1/p - 1$.*

*Proof.* The function $\frac{\mathrm{d}a^{-1}\sharp\mu^\delta}{\mathrm{d}\mu_U} : U \to \mathbb{R}$ can be expressed as the composition of the maps

$$\boldsymbol{y} \mapsto u(\boldsymbol{y}), \quad u \mapsto \frac{1}{2}(\delta - \mathcal{O}(u))^\top\Gamma^{-1}(\delta - \mathcal{O}(u)), \quad a \mapsto \exp(-a). \tag{15.40}$$

The first map is $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphic, the second map is a holomorphic mapping from $\mathcal{X}$ to $\mathbb{C}$, and the third map is holomorphic from $\mathbb{C}$ to $\mathbb{C}$. The composition is $(\boldsymbol{b}, \varepsilon, \mathbb{R})$-holomorphic. The rest of the statement follows by Theorem 15.4.9. □

In case the number of parameters $N \in \mathbb{N}$ is finite, *exponential convergence rates of ReLU DNN approximations* follow with [62, Theorem 3.6] but with the rate of convergence and other constants in the error bound depending on $N$.

For the approximation of the posterior expectation $Y \to Z : \delta \mapsto \mathbb{E}[Q \circ u|\delta]$, holomorphy of the posterior density implies holomorphy of the posterior expectation but without control on the size of the domain of holomorphy. Thus [62, Theorem 3.6] gives exponential convergence with rate $C\exp(-b\mathcal{N}^{1/(K+1)})$, with possibly very small $b > 0$, in terms of the NN size $\mathcal{N}$. We remark that holomorphy of the data-to-QoI map is valid even for nonholomorphic input-to-response maps in the operator equation [35]. In [35], this was exploited by considering a *rational approximation of the Bayesian estimate* based on

$$\delta \mapsto \mathbb{E}[Q \circ u|\delta] = \int_{\tilde{X}} Q(u(a))\frac{1}{Z(\delta)}\exp(-\Phi(a;\delta))\mathrm{d}\mu_0(a) =: Z'(\delta)/Z(\delta),$$

where $Z, Z'$ are entire functions of $\delta$, i. e., they admit a holomorphic extension to $\mathbb{C}^K$. With that argument, convergence rates of the form $C\exp(-b\mathcal{N}^{1/(K+1)})$ with arbitrarily large $b > 0$ were obtained.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

## 15.6.2 Posterior concentration

We consider the DNN expression of posterior densities in Bayesian inverse problems when the posterior density concentrates near a single point, the so-called *maximum a posteriori point (MAP point)*, at which the posterior density attains its maximum.

We consider in particular the case in which the posterior density exists, is unimodal, and attains its global maximum at the MAP point. In the mentioned scaling regimes, in the vicinity of the MAP point, the Bayesian posterior density is close to a Gaussian distribution with covariance matrix $\Gamma$, which arises in either the small noise or the large data limits; cf., e. g., [42, 71]. We therefore study the behavior of the DNN expression rate bounds as $\Gamma \downarrow 0$. This limit applies to the situation of decreasing observation noise $\eta$ or of increasing observation size $\dim(Y)$.

The results in Section 15.6.1 hold for all symmetric positive definite covariance matrices $\Gamma$, but constants depend on $\Gamma$ and may tend to infinity as $\Gamma \downarrow 0$. However, the concentration can be exploited for the approximation of the posterior density. As an example, we consider an inverse problem with $N < \infty$ parameters, with a holomorphic forward map $[-1,1]^N \to \mathcal{X} : \boldsymbol{y} \to u(\boldsymbol{y})$, a linear observation functional $\mathcal{O} : \mathcal{X} \to Y$, and a finite observation size $K := \dim(Y) < \infty$. In [70, Theorem 4.1], in case of a nondegenerate Hessian $\Phi_{\boldsymbol{y},\boldsymbol{y}}$, it was shown that after a $\Gamma$-dependent affine transformation, the posterior density is analytic with polyradii of analyticity independent of $\Gamma$. Hence by [62, Theorem 3.6] NN approximations of the posterior density converge exponentially (albeit with constants depending exponentially on $N$).

Moreover, in [71, Appendix], it was shown that under suitable conditions, a Gaussian distribution approximates the posterior density up to first order in $\Gamma$. This allows us to overcome the curse of dimensionality in terms of $N$ for the unnormalized posterior density by exploiting the radial symmetry of the Gaussian density function. By [60, Theorem 6.7] the Gaussian density function can be approximated by ReLU NNs with the network size growing polylogarithmically with the error and the corresponding constants increasing at most quadratically in $N$. Thus there is no curse of dimensionality for the approximation of the unnormalized posterior density when it concentrates near one point. Note that this ignores the consistency error of the posterior density with respect to this Gaussian approximation to the true posterior density. If the posterior concentrates near multiple well-separated points and if it is close to a Gaussian near each of the points, then it can be approximated at the same rate by a sum of (localized) Gaussians.

The next proposition gives an approximation result for unnormalized Gaussian densities. We refer to [61, Appendix A.5] for a proof.

**Proposition 15.6.2.** *For $N \in \mathbb{N}$, let $A : \mathbb{R}^N \to \mathbb{R}^N$ be a bijective linear map. For $\boldsymbol{x} \in \mathbb{R}^N$, set $\mathfrak{g}(\boldsymbol{x}) := \exp(-\frac{1}{2}\|A\boldsymbol{x}\|_2^2)$.*

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

*Then there exists $C > 0$ independent of $A$ and $N$ such that for every $\varepsilon \in (0,1)$, there exists a ReLU NN $\Phi_\varepsilon^{\mathfrak{g}}$ satisfying*

$$\begin{aligned}\left\|\mathfrak{g} - \Phi_\varepsilon^{\mathfrak{g}}\right\|_{L^\infty(\mathbb{R}^N)} &\leq C\varepsilon = C\varepsilon\|\mathfrak{g}\|_{L^\infty(\mathbb{R}^N)},\\ \operatorname{depth}(\Phi_\varepsilon^{\mathfrak{g}}) &\leq C(\log(N)(1+\log(N/\varepsilon)) + 1 + \log(1/\varepsilon)\log\log(1/\varepsilon)),\\ \operatorname{size}(\Phi_\varepsilon^{\mathfrak{g}}) &\leq C\big((1+\log(1/\varepsilon))^2 + N\log(1/\varepsilon) + N^2\big).\end{aligned}$$

**Remark 15.6.3.** The term $CN^2$ in the bound on the network size follows from bounding the number of nonzero coefficients in the linear map $A$ by $N^2$. If $A$ has at most $CN$ nonzero coefficients, then the network size is of the order $N\log(N)$.

Densities of the type $\mathfrak{g}(\boldsymbol{x}) = \exp(-\frac{1}{2}\|A(\boldsymbol{x})\|_2^2)$ need to be normalized to become probability densities on $[-1,1]^N$. We now discuss an example to show the effect of the normalization constant on the approximation result when the density concentrates.

Fix an observation noise covariance $\Gamma \in \mathbb{R}^{N\times N}$, which symmetric positive definite, and for $n \in \mathbb{N}$, set $\Gamma_n := \Gamma/n$ and $\tilde{\mathfrak{g}}_n(\boldsymbol{x}) = \exp(-\frac{1}{2}\|\Gamma_n^{-1/2}\boldsymbol{x}\|_2^2)$ for $\boldsymbol{x} \in [-1,1]^N$. Given $\delta \in [-1,1]^N$, note that as $n \to \infty$, the unnormalized density $\tilde{\mathfrak{g}}_n(\boldsymbol{x}-\delta)$ concentrates around $\delta \in [-1,1]^N$. For any $n \geq 1$, using the change of variables $\boldsymbol{y} = \sqrt{n}\boldsymbol{x}$, we bound the normalization constant from below:

$$\begin{aligned}\int_{[-1,1]^N} \tilde{\mathfrak{g}}_n(\boldsymbol{x}-\delta)\mathrm{d}\boldsymbol{x} &= \int_{[-1,1]^N} \exp\Big(-\frac{1}{2}\big\|\sqrt{n}\Gamma^{-1/2}(\boldsymbol{x}-\delta)\big\|_2^2\Big)\mathrm{d}\boldsymbol{x}\\ &= n^{-N/2}\int_{[-\sqrt{n},\sqrt{n}]^N} \exp\Big(-\frac{1}{2}\big\|\Gamma^{-1/2}(\boldsymbol{y}-\sqrt{n}\delta)\big\|_2^2\Big)\mathrm{d}\boldsymbol{y}\\ &\geq n^{-N/2}\inf_{\tilde{\delta}\in[-1,1]^N}\int_{[-1,1]^N} \exp\Big(-\frac{1}{2}\big\|\Gamma^{-1/2}(\boldsymbol{y}-\tilde{\delta})\big\|_2^2\Big)\mathrm{d}\boldsymbol{y}\\ &= n^{-N/2}C_0,\end{aligned}$$

where $C_0(\Gamma,N) > 0$ denotes the infimum in the second to last line, and where we used $\sqrt{n}\delta \in [-\sqrt{n},\sqrt{n}]^N$.

Denote $Z_n(\delta) := \int_{[-1,1]^N}\tilde{\mathfrak{g}}_n(\boldsymbol{x}-\delta)\mathrm{d}\boldsymbol{x} \geq C_0 n^{-N/2}$. Then by Proposition 15.6.2 the normalized density $\mathfrak{g}_n(\boldsymbol{x}-\delta) := \tilde{\mathfrak{g}}_n(\boldsymbol{x}-\delta)/Z_n(\delta) \leq C_0^{-1}n^{N/2}\tilde{\mathfrak{g}}_n(\boldsymbol{x}-\delta)$ can be uniformly approximated on $[-1,1]^N$ to accuracy $\varepsilon > 0$ with a ReLU network $\Phi_\varepsilon^{\mathfrak{g}_n}$ of size and depth bounded as follows: for $C(\Gamma,N) > 0$,

$$\begin{aligned}\operatorname{depth}(\Phi_\varepsilon^{\mathfrak{g}_n}) &\leq C(1+(\log(1/\varepsilon)+(1+\log(n)))\log(\log(1/\varepsilon)+(1+\log(n)))),\\ \operatorname{size}(\Phi_\varepsilon^{\mathfrak{g}_n}) &\leq C\big((1+\log(1/\varepsilon))^2 + \log(1/\varepsilon)(1+\log_2(n)) + (1+\log_2(n))^2\big).\end{aligned}$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

### 15.6.3 Posterior consistency

In Section 15.6.1, we proved $L^\infty(U)$-bounds on the approximation of the posterior density with NNs. Up to a constant, this immediately yields the same bounds for the Hellinger and total-variation distances of the corresponding (normalized) Bayesian posterior measures, as we show next.

Let $\lambda$ be the Lebesgue measure on $[-1,1]$, and denote again by $\mu_U := \bigotimes_{j\in\mathbb{N}} \frac{\lambda}{2}$ the uniform probability measure on $U = [-1,1]^{\mathbb{N}}$ equipped with the product sigma-algebra. Let $\mu \ll \mu_U$ and $\nu \ll \mu_U$ be two measures on $U$ with Radon–Nikodym derivatives $\frac{\mathrm{d}\mu}{\mathrm{d}\mu_U} =: \pi_\mu : U \to \mathbb{R}$ and $\frac{\mathrm{d}\nu}{\mathrm{d}\mu_U} =: \pi_\nu : U \to \mathbb{R}$. Recall that the Hellinger distance (which we use here also for nonprobability measures) is defined as

$$d_H(\mu,\nu) = \left(\frac{1}{2}\int_U \big(\sqrt{\pi_\mu(\boldsymbol{y})} - \sqrt{\pi_\nu(\boldsymbol{y})}\big)^2 \mathrm{d}\mu_U(\boldsymbol{y})\right)^{1/2} = \frac{1}{\sqrt{2}}\|\sqrt{\pi_\mu} - \sqrt{\pi_\nu}\|_{L^2(U,\mu_U)}.$$

The total-variation distance is defined as

$$d_{\mathrm{TV}}(\mu,\nu) = \sup_B |\mu(B) - \nu(B)| \le \int_U |\pi_\mu(\boldsymbol{y}) - \pi_\nu(\boldsymbol{y})| \mathrm{d}\mu_U(\boldsymbol{y}) = \|\pi_\mu - \pi_\nu\|_{L^1(U,\mu_U)},$$

where the supremum is taken over all measurable $B \subseteq U$. Thus

$$d_{\mathrm{TV}}(\mu,\nu) \le \|\pi_\mu - \pi_\nu\|_{L^\infty(U,\mu_U)}.$$

Since $|\sqrt{x} - \sqrt{y}| = \frac{|x-y|}{|\sqrt{x}+\sqrt{y}|}$ for all $x, y \ge 0$,

$$d_H(\mu,\nu) = \frac{1}{\sqrt{2}}\|\sqrt{\pi_\mu} - \sqrt{\pi_\nu}\|_{L^2(U,\mu_U)} \le \frac{\|\pi_\mu - \pi_\nu\|_{L^\infty(U,\mu_U)}}{\sqrt{2}\inf_{\boldsymbol{y}\in U}(\sqrt{\pi_\mu(\boldsymbol{y})} + \sqrt{\pi_\nu(\boldsymbol{y})})}.$$

Denote by $\overline{\mu} = \frac{\mu}{\mu(U)}$ and $\overline{\nu} = \frac{\nu}{\nu(U)}$ the normalized measures and by $\overline{\pi}_\mu$ and $\overline{\pi}_\nu$ the corresponding densities (which are probability densities with respect to $\mu_U$). Then for all $\boldsymbol{y} \in U$,

$$\begin{aligned}|\overline{\pi}_\mu(\boldsymbol{y}) - \overline{\pi}_\nu(\boldsymbol{y})| &= \left|\frac{\pi_\mu(\boldsymbol{y})}{\mu(U)} - \frac{\pi_\nu(\boldsymbol{y})}{\nu(U)}\right| \\ &\le \frac{|\pi_\mu(\boldsymbol{y})\nu(U) - \pi_\nu(\boldsymbol{y})\nu(U)| + |\pi_\nu(\boldsymbol{y})\nu(U) - \pi_\nu(\boldsymbol{y})\mu(U)|}{\mu(U)\nu(U)}.\end{aligned}$$

Using $|\mu(U) - \nu(U)| \le \|\pi_\mu - \pi_\nu\|_{L^1(U,\mu_U)}$, we obtain, for all $\boldsymbol{y} \in U$,

$$|\overline{\pi}_\mu(\boldsymbol{y}) - \overline{\pi}_\nu(\boldsymbol{y})| \le \frac{\|\pi_\mu - \pi_\nu\|_{L^\infty(U,\mu_U)}\nu(U) + \|\pi_\nu\|_{L^\infty(U,\mu_U)}\|\pi_\mu - \pi_\nu\|_{L^\infty(U,\mu_U)}}{\nu(U)\mu(U)}.$$

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

By symmetry this implies

$$d_{\mathrm{TV}}(\overline{\mu},\overline{\nu}) \leq \|\pi_\mu - \pi_\nu\|_{L^\infty(U,\mu_U)} \min\left(\frac{\nu(U)+\|\pi_\nu\|_{L^\infty(U,\mu_U)}}{\nu(U)\mu(U)}, \frac{\mu(U)+\|\pi_\mu\|_{L^\infty(U,\mu_U)}}{\nu(U)\mu(U)}\right), \quad (15.41a)$$

and similarly, as before,

$$d_H(\overline{\mu},\overline{\nu}) \leq \|\pi_\mu - \pi_\nu\|_{L^\infty(U,\mu_U)} \frac{\min(\frac{\nu(U)+\|\pi_\nu\|_{L^\infty(U,\mu_U)}}{\nu(U)\mu(U)}, \frac{\mu(U)+\|\pi_\mu\|_{L^\infty(U,\mu_U)}}{\nu(U)\mu(U)})}{\sqrt{2}\inf_{\boldsymbol{y}\in U}(\sqrt{\bar{\pi}_\mu}(\boldsymbol{y}) + \sqrt{\bar{\pi}_\nu}(\boldsymbol{y}))}. \quad (15.41b)$$

**Proposition 15.6.4.** *Consider the setting of Corollary* 15.6.1. *Then for every* $\tau \in (0,1)$, *there exists a* $\sigma_1$*-NN* $f_\tau : U \to [0,\infty)$ *(with input variables* $(y_j)_{j\in S_{\Lambda_\tau}}$*) such that with* $\Lambda_\tau$ *as in Theorem* 15.3.7

$$\begin{aligned} \mathrm{size}(f_\tau) &\leq C(1+|\Lambda_\tau|\cdot\log|\Lambda_\tau|\cdot\log\log|\Lambda_\tau|), \\ \mathrm{depth}(f_\tau) &\leq C(1+\log|\Lambda_\tau|\cdot\log\log|\Lambda_\tau|), \end{aligned} \quad (15.42)$$

*and the measure* $\nu_\tau$ *on* $U$ *with density* $f_\tau = \frac{\mathrm{d}\nu_\tau}{\mathrm{d}\mu_U}$ *satisfies*

$$d_H(a^{-1}\sharp\mu^\delta, \bar{\nu}_\tau) \leq C|\Lambda_\tau|^{-\frac{1}{p}+1}, \quad (15.43)$$

*and the same bound holds with respect to* $d_{\mathrm{TV}}$.

*Proof.* By Corollary 15.6.1 and Theorem 15.4.9 there exists a $\sigma_1$-NN $\tilde{f}_\tau : U \to \mathbb{R}$ satisfying (15.42) such that with $f(\boldsymbol{y}) := \frac{\mathrm{d}a^{-1}\sharp\mu^\delta}{\mathrm{d}\mu_U} = \frac{1}{Z(\delta)}\exp(-\frac{1}{2}\|\delta - \mathcal{O}(u(\boldsymbol{y}))\|_\Gamma^2)$, where $u$ is $(\boldsymbol{b},\varepsilon)$-holomorphic, we have

$$\|f - \tilde{f}_\tau\|_{L^\infty(U)} \leq C|\Lambda_\tau|^{-\frac{1}{p}+1}. \quad (15.44)$$

Let $f_\tau := \sigma_1(\tilde{f}_\tau)$. Then $f_\tau : U \to [0,\infty)$, and the bound (15.44) remains true for $f_\tau$ since $f(\boldsymbol{y}) \geq 0$ for all $\boldsymbol{y} \in U$.

Since any $(\boldsymbol{b},\varepsilon)$-holomorphic function is continuous on $U$ and because $f(\boldsymbol{y}) > 0$ for all $\boldsymbol{y} \in U$, we have $\inf_{\boldsymbol{y}\in U} f(\boldsymbol{y}) > 0$ and $\sup_{\boldsymbol{y}\in U} f(\boldsymbol{y}) < \infty$. Thus (15.41) implies (15.43) for $d_{\mathrm{TV}}$ and $d_H$. □

## 15.7 Conclusions and further directions

In this paper, we presented dimension-independent expression rates for the approximation of infinite-parametric functions occurring in forward and inverse UQ by deep neural networks. Our results are based on multilevel gPC expansions and generalize the statements of [74] in that they do not require analytic extensions of the target function to complex polydiscs, but merely to complex polyellipses. Additionally, whereas

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

for $\mathcal{X}$-valued functions, [74] only treated the case of $\mathcal{X} = H^1([0,1])$, here we considered $\mathcal{X} = W^{1,q}(\mathrm{D})$ with D being, for example, a bounded polytope. It was shown that our theory also comprises analyticity of parametric maps in scales of corner-weighted Sobolev spaces in D, allowing us to retain optimal convergence rates of FEM in the presence of corner singularities of the PDE solution. These generalizations allow us to treat much wider problem classes, comprising, for example, a forward operator mapping inputs to the solution of the parametric (nonlinear) Navier–Stokes equations [17]. Another instance includes domain uncertainty, which typically does not yield forward operators with holomorphic parameter dependence on polydiscs; see, e. g., [36].

As one possible application of our results, we treated in more detail the approximation of posterior densities in Bayesian inference. Having cheaply evaluable surrogates of this density (in the form of a DNN) can be a powerful tool, as any inference technique can require thousands of evaluations of the posterior density. On top of that, in case of MCMC, arguably the most widely used inference algorithm, these evaluations are inherently sequential and not parallel. Each such evaluation requires a (time-consuming, approximate) computation of a PDE solution, which can render MCMC infeasible in practice. Variational inference, on the other hand, where sampling from the posterior is replaced by an optimization problem, does not necessarily require sequential computation of (approximate) PDE solutions; however, it still demands a high number of evaluations of the posterior, which may be significantly sped up if this posterior is replaced by a cheap surrogate. We refer, for example, to transport-based methods such as [50].

As already indicated in the introduction of the present paper, the idea of using DNNs for expressing the input-to-response map (i. e., the "forward" map) for PDE models has been proposed repeatedly in recent years. The motivation for this is the nonlinearity of such maps, even for linear PDEs, and the often high regularity (e. g., holomorphy) of such maps. Here DNNs are a computational tool alongside other reduction methods, such as reduced basis (RB) or model order reduction (MOR) methods. Indeed, in [43, Remark 4.6], it is suggested that under the provision that reduced bases for a compact solution manifold of a linear elliptic parametric PDE admit an efficient DNN expression, so does the input-to-solution map of this PDE. The abstract, Lipschitz dependence result, Theorem 15.2.8 (which is [21, Theorem 18]), will imply with the present results and the DNN expression results of RB/MOR approximations for forward PDE problems as developed in [43] analogous results also for the corresponding Bayesian inverse problems considered in the present paper. MOR and RB approaches can be developed along the lines of [11], where BIP subject RB/MOR approximation of the forward input-to-response maps were considered in conjunction with Bayesian inverse problems of the type considered here. Should reduced bases admit good DNN expression rates, the analysis of [11] would imply with the present results the corresponding improved DNN expression rates along the lines of [43].

We remark that the DNN expression rate bounds for the posterior densities are obtained from DNN reapproximation of gPC surrogates. DNN expression rate bounds

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

follow from the corresponding approximation rates of $N$-term truncated gPC expansions. These, in turn, are based on gPC coefficient estimates, which were obtained as, e. g., in [74] by analytic continuation of parametric solution families into the complex domain. Analytic continuation can be avoided if, instead, real-variable induction arguments for bounding derivatives of parametric solutions are employed. We refer to [30] for forward UQ in an elliptic control problem and to [33, Section 7] for a proof of derivative bounds for the Bayesian posterior with Gaussian prior. As in [74], the present DNN expression rate analysis relies on "intermediate" polynomial chaos approximations of the posterior density, assuming a prior given by the uniform probability measure on $U = [-1,1]^{\mathbb{N}}$. However, the emulation of the posterior density by DNNs can leverage the *compositional structure* of DNNs to accommodate changes of (prior) probability, with essentially the same expression rates, as long as the changes of measure can be emulated efficiently by DNNs. This may include nonanalytic/nonholomorphic densities. We refer to [62, Section 4.3.5] for an example.

We also showed in Section 15.6.2 that ReLU DNN expression rates are either independent of or depend only logarithmically on concentration in the posterior density, provided that the concentration happens only in a finite number of "informed" variables, and the posterior density is of "MAP" type, in particular (locally), unimodal. Although important, this is only a rather particular case in applications, where oftentimes posterior concentration occurs along smooth submanifolds. In such cases, ReLU DNNs can also be expected to exhibit robust expression rates according to the expression rate bounds in [64, Section 5]. Details are to be developed elsewhere.

# Bibliography

[1] R. Andreev and C. Schwab. Sparse tensor approximation of parametric eigenvalue problems. In *Numerical Analysis of Multiscale Problems*, volume 83 of Lecture Notes in Computational Science and Engineering, pages 203–241. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[2] I. Babuška, R. B. Kellogg, and J. Pitkäranta. Direct and inverse error estimates for finite elements with mesh refinements. *Numer. Math.*, 33(4):447–471, 1979.

[3] M. Bachmayr, A. Cohen, D. Dũng, and C. Schwab. Fully discrete approximation of parametric and stochastic elliptic PDEs. *SIAM J. Numer. Anal.*, 55(5):2151–2186, 2017.

[4] M. Bachmayr, A. Cohen, and G. Migliorati. Sparse polynomial approximation of parametric elliptic PDEs. Part I: affine coefficients. *ESAIM: Math. Model. Numer. Anal.*, 51(1):321–339, 2017.

[5] C. Băcuţă, H. Li, and V. Nistor. Differential operators on domains with conical points: precise uniform regularity estimates. *Rev. Roum. Math. Pures Appl.*, 62(3):383–411, 2017.

[6] C. Băcuţă, V. Nistor, and L. T. Zikatanov. Improving the rate of convergence of 'high order finite elements' on polygons and domains with cusps. *Numer. Math.*, 100(2):165–184, 2005.

[7] A. R. Barron. Complexity regularization with application to artificial neural networks. In *Nonparametric Functional Estimation and Related Topics (Spetses, 1990)*, volume 335 of NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., pages 561–576. Kluwer Acad. Publ., Dordrecht, 1991.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[8] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory*, 39(3):930–945, 1993.
[9] J. Berg and K. Nyström. Neural network augmented inverse problems for PDEs. *arXiv preprint* arXiv:1712.09685, 2017.
[10] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen. Optimal approximation with sparsely connected deep neural networks. *SIAM J. Math. Data Sci.*, 1(1):8–45, 2019.
[11] P. Chen and C. Schwab. Sparse-grid, reduced-basis Bayesian inversion: nonaffine-parametric nonlinear equations. *J. Comput. Phys.*, 316:470–503, 2016.
[12] A. Chkifa, A. Cohen, and C. Schwab. High-dimensional adaptive sparse polynomial interpolation and applications to parametric PDEs. *Found. Comput. Math.*, 14(4):601–633, 2013.
[13] A. Cohen, A. Chkifa, and C. Schwab. Breaking the curse of dimensionality in sparse polynomial approximation of parametric PDEs. *J. Math. Pures Appl.*, 103(2):400–428, 2015.
[14] A. Cohen and R. DeVore. Approximation of high-dimensional parametric PDEs. *Acta Numer.*, 24:1–159, 2015.
[15] A. Cohen, R. DeVore, and C. Schwab. Convergence rates of best *N*-term Galerkin approximations for a class of elliptic sPDEs. *Found. Comput. Math.*, 10(6):615–646, 2010.
[16] A. Cohen, R. Devore, and C. Schwab. Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's. *Anal. Appl. (Singap.)*, 9(1):11–47, 2011.
[17] A. Cohen, C. Schwab, and J. Zech. Shape holomorphy of the stationary Navier–Stokes equations. *SIAM J. Math. Anal.*, 50(2):1720–1752, 2018.
[18] N. Cohen, O. Sharir, A. Shashua, On the expressive power of deep learning: a tensor analysis. In *Proc. of 29th Ann. Conf. Learning Theory*, pages 698–728, 2016. arXiv:1509.05009v3.
[19] S. Dahlke, M. Hansen, C. Schneider, and W. Sickel. Properties of Kondratiev spaces. *arXiv preprint* arXiv:1911.01962, 2019.
[20] M. Dashti, S. Harris, and A. Stuart. Besov priors for Bayesian inverse problems. *Inverse Probl. Imaging*, 6(2):183–200, 2012.
[21] M. Dashti and A. M. Stuart. The Bayesian approach to inverse problems. In *Handbook of Uncertainty Quantification. Vol. 1, 2, 3*, pages 311–428. Springer, Cham, 2017.
[22] J. Daws Jr. and C. G. Webster. A polynomial-based approach for architectural design and learning with deep neural networks. *arXiv preprint* arXiv:1905.10457, 2019.
[23] P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *Ann. Stat.*, 14(1):1–67, 1986. With a discussion and a rejoinder by the authors.
[24] J. Dick, R. N. Gantner, Q. T. Le Gia, and C. Schwab. Multilevel higher-order quasi-Monte Carlo Bayesian estimation. *Math. Models Methods Appl. Sci.*, 27(5):953–995, 2017.
[25] D. Dũng. Linear collective collocation and Galerkin approximations for parametric and stochastic elliptic PDEs. *arXiv preprint* arXiv:1511.03377, 2015.
[26] D. Elbrächter, P. Grohs, A. Jentzen, and C. Schwab. DNN expression rate analysis of high-dimensional PDEs: application to option pricing. *Constr. Approx.*, 2021. https://doi.org/10.1007/s00365-021-09541-6.
[27] K.-I. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Netw.*, 2(3):183–192, 1989.
[28] A. D. Gilbert, I. G. Graham, F. Y. Kuo, R. Scheichl, and I. H. Sloan. Analysis of quasi-Monte Carlo methods for elliptic eigenvalue problems with stochastic coefficients. *Numer. Math.*, 142(4):863–915, 2019.
[29] P. Grohs, T. Wiatowski, and H. Bölcskei. Deep convolutional neural networks on cartoon functions. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 1163–1167, 2016.
[30] P. A. Guth, V. Kaarnioja, F. Y. Kuo, C. Schillings, and I. H. Sloan. A quasi-Monte Carlo method for optimal control under uncertainty. *SIAM/ASA J. Uncertain. Quantificat.*, 9(2):354–383, 2021.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[31] J. He, L. Li, J. Xu, and C. Zheng. ReLU deep neural networks and linear finite elements. *J. Comput. Math.*, 38:502–527, 2020.
[32] F. Henríquez and C. Schwab. Shape holomorphy of the Calderón projector for the Laplacian in $\mathbb{R}^2$. *Integral Equ. Oper. Theory*, 93(4):43, 2021.
[33] L. Herrmann, M. Keller, and C. Schwab. Quasi-Monte Carlo Bayesian estimation under Besov priors in elliptic inverse problems. *Math. Comput.*, 90:1831–1860, 2021.
[34] L. Herrmann and C. Schwab. Multilevel quasi-Monte Carlo uncertainty quantification for advection–diffusion–reaction. In *Monte Carlo and Quasi-Monte Carlo Methods*, volume 324 of Springer Proceedings in Mathematics and Statistics, pages 31–67. Springer, Cham, 2020.
[35] L. Herrmann, C. Schwab, and J. Zech. Deep neural network expression of posterior expectations in Bayesian PDE inversion. *Inverse Probl.*, 36(12):125011, 2020.
[36] R. Hiptmair, L. Scarabosio, C. Schillings, and C. Schwab. Large deformation shape uncertainty quantification in acoustic scattering. *Adv. Comput. Math.*, 44(5):1475–1518, 2018.
[37] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Netw.*, 4(2):251–257, 1991.
[38] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Netw.*, 2(5):359–366, 1989.
[39] B. Hosseini and N. Nigam. Well-posed Bayesian inverse problems: priors with exponential tails. *SIAM/ASA J. Uncertain. Quantificat.*, 5(1):436–465, 2017.
[40] T. Y. Hou, K. C. Lam, P. Zhang, and S. Zhang. Solving Bayesian inverse problems from the perspective of deep generative networks. *Comput. Mech.*, 64(2):395–408, 2019.
[41] C. Jerez-Hanckes, C. Schwab, and J. Zech. Electromagnetic wave scattering by random surfaces: shape holomorphy. *Math. Models Methods Appl. Sci.*, 27(12):2229–2259, 2017.
[42] B. T. Knapik, A. W. van der Vaart, and J. H. van Zanten. Bayesian inverse problems with Gaussian priors. *Ann. Stat.*, 39(5):2626–2657, 2011.
[43] G. Kutyniok, P. Petersen, M. Raslan, and R. Schneider. A theoretical analysis of deep neural networks and parametric PDEs. *Constr. Approx.*, 2021. https://doi.org/10.1007/s00365-021-09551-4.
[44] M. Lassas, E. Saksman, and S. Siltanen. Discretization-invariant Bayesian inversion and Besov space priors. *Inverse Probl. Imaging*, 3(1):87–122, 2009.
[45] B. Li, S. Tang, and H. Yu. Better approximations of high dimensional smooth functions by deep neural networks with rectified power units. *Commun. Comput. Phys.*, 27(2):379–411, 2020.
[46] H. Li, A. Mazzucato, and V. Nistor. Analysis of the finite element method for transmission/mixed boundary value problems on general polygonal domains. *Electron. Trans. Numer. Anal.*, 37:41–69, 2010.
[47] L. Liang, F. Kong, C. Martin, T. Pham, Q. Wang, J. Duncan, and W. Sun. Machine learning-based 3-D geometry reconstruction and modeling of aortic valve deformation using 3-D computed tomography images. *Int. J. Numer. Methods Biomed. Eng.*, 33(5):e2827, 13, 2017.
[48] L. Liang, M. Liu, C. Martin, and W. Sun. A deep learning approach to estimate stress distribution: a fast and accurate surrogate of finite-element analysis. *J. R. Soc. Interface*, 15:20170844, 2018.
[49] S. Liang and R. Srikant. Why deep neural networks for function approximation? In *Proc. of ICLR 2017*, 2017. arXiv:1610.04161.
[50] Y. Marzouk, T. Moselhy, M. Parno, and A. Spantini. Sampling via measure transport: an introduction. In *Handbook of Uncertainty Quantification. Vol. 1, 2, 3*, pages 785–825. Springer, Cham, 2017.
[51] V. Maz'ya and J. Rossmann. *Elliptic Equations in Polyhedral Domains*, volume 162 of Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2010.
[52] H. N Mhaskar. Neural networks for localized approximation of real functions. In *Neural Networks for Signal Processing III – Proceedings of the 1993 IEEE-SP Workshop*, pages 190–196. IEEE, 1993.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[53] H. N. Mhaskar. Approximation properties of a multilayered feedforward artificial neural network. *Adv. Comput. Math.*, 1(1):61–80, Feb 1993.
[54] H. N. Mhaskar and T. Poggio. Deep vs. shallow networks: an approximation theory perspective. *Anal. Appl. (Singap.)*, 14(6):829–848, 2016.
[55] J. R. Munkres. *Topology*, 2nd edition. Prentice Hall, Inc., Upper Saddle River, NJ, 2000.
[56] G. A. Muñoz, Y. Sarantopoulos, and A. Tonge. Complexifications of real Banach spaces, polynomials and multilinear maps. *Stud. Math.*, 134(1):1–33, 1999.
[57] F. Nobile, R. Tempone, and C. G. Webster. An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM J. Numer. Anal.*, 46(5):2411–2442, 2008.
[58] R. H. Nochetto and A. Veeser. Primer of adaptive finite element methods. In *Multiscale and Adaptivity: Modeling, Numerics and Applications*, volume 2040 of Lecture Notes in Math., pages 125–225. Springer, Heidelberg, 2012.
[59] F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, editors. *NIST Handbook of Mathematical Functions*. US Department of Commerce, National Institute of Standards and Technology, Washington, DC; Cambridge University Press, Cambridge, 2010. With 1 CD-ROM (Windows, Macintosh and UNIX).
[60] J. A. A. Opschoor, P. C. Petersen, and C. Schwab. Deep ReLU networks and high-order finite element methods. *Anal. Appl.*, 18(05):715–770, 2020.
[61] J. A. A. Opschoor, C. Schwab, and J. Zech. Deep learning in high dimension: ReLU neural network expression for Bayesian PDE inversion (extended version). Technical Report 2020-47, Seminar for Applied Mathematics, ETH Zürich, Switzerland, 2022.
[62] J. A. A. Opschoor, C. Schwab, and J. Zech. Exponential ReLU DNN expression of holomorphic maps in high dimension. *Constr. Approx.*, 2021. https://doi.org/10.1007/s00365-021-09542-5.
[63] G. Pang, L. Lu, and G. E. Karniadakis. fPINNs: fractional physics-informed neural networks. *SIAM J. Sci. Comput.*, 41(4):A2603–A2626, 2019.
[64] P. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Netw.*, 108:296–330, 2018.
[65] A. Pinkus. Approximation theory of the MLP model in neural networks. In *Acta Numerica, 1999*, volume 8 of Acta Numer., pages 143–195. Cambridge Univeristy Press, Cambridge, 1999.
[66] A. Quarteroni and G. Rozza. Numerical solution of parametrized Navier–Stokes equations by reduced basis methods. *Numer. Methods Partial Differ. Equ.*, 23(4):923–948, 2007.
[67] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.*, 378:686–707, 2019.
[68] S. Reich and C. Cotter. *Probabilistic forecasting and Bayesian data assimilation*. Cambridge University Press, New York, 2015.
[69] D. Rolnik and M. Tegmark. The power of deeper networks for expressing natural functions. In *International Conference on Learning Representations*, 2018.
[70] C. Schillings and C. Schwab. Sparsity in Bayesian inversion of parametric operator equations. *Inverse Probl.*, 30(6):065007, 2014.
[71] C. Schillings and C. Schwab. Scaling limits in computational Bayesian inversion. *ESAIM: M2AN*, 50(6):1825–1856, 2016.
[72] C. Schwab and A. M. Stuart. Sparse deterministic approximation of Bayesian inverse problems. *Inverse Probl.*, 28(4):045003, 32, 2012.
[73] C. Schwab and R. A. Todor. Convergence rates of sparse chaos approximations of elliptic problems with stochastic coefficients. *IMA J. Numer. Anal.*, 44:232–261, 2007.
[74] C. Schwab and J. Zech. Deep learning in high dimension: neural network expression rates for generalized polynomial chaos expansions in UQ. *Anal. Appl. (Singap.)*, 17(1):19–55, 2019.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

[75] J. Sirignano and K. Spiliopoulos. DGM: a deep learning algorithm for solving partial differential equations. *J. Comput. Phys.*, 375:1339–1364, 2018.
[76] A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numer.*, 19:451–559, 2010.
[77] T. J. Sullivan. Well-posed Bayesian inverse problems and heavy-tailed stable quasi-Banach space priors. *Inverse Probl. Imaging*, 11(5):857–874, 2017.
[78] J. M Tarela and M. V Martinez. Region configurations for realizability of lattice piecewise-linear models. *Math. Comput. Model.*, 30(11–12):17–27, 1999.
[79] R. K. Tripathy and I. Bilionis. Deep UQ: learning deep neural network surrogate models for high dimensional uncertainty quantification. *J. Comput. Phys.*, 375:565–588, 2018.
[80] D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Netw.*, 94:103–114, 2017.
[81] J. Zech. Sparse-grid approximation of high-dimensional parametric PDEs. PhD thesis, 2018. Dissertation 25683, ETH Zürich.
[82] J. Zech, D. Dung, and C. Schwab. Multilevel approximation of parametric and stochastic PDEs. *M3AS*, 29(9):1753–1817, 2019.
[83] J. Zech and C. Schwab. Convergence rates of high dimensional Smolyak quadrature. *ESAIM: Math. Model. Numer. Anal.*, 54(4):1259–1307, 2020.
[84] D. Zhang, L. Lu, L. Guo, and G. E. Karniadakis. Quantifying total uncertainty in physics-informed neural networks for solving forward and inverse stochastic problems. *J. Comput. Phys.*, 397:108850, 19, 2019.

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

# Index

a priori error estimates 158
adaptive finite elements 217
adaptive reduced basis 17
adaptivity 184

Bayesian inverse problems 420
boundary control 60
branched transport 190

Caginalp 367
Cahn–Hilliard 367
calibration 211
continuous space-time finite element methods 175
convexification 199

damped Newton method 288
discrete sparsity 114
double-well potential 368

ensemble Kalman filter 402
ensemble Kalman inversion 402
epidemic vector control 74
exact controllability 33

feedback controller 376
feedback law 60
finite element approximation 35
finite element method 147
functional lifting 197

gas pipeline flow 59
generalized polynomial chaos 422

higher order models 262
homogenization 237

initial measure control 133
inverse problems 395

least squares 287
logarithmic potential 368

minimal time 81
mixed formulation 35
model order reduction 16
Mumford–Shah problem 201

Navier–Stokes equations 286
necessary optimality conditions 152
neural networks 400
non-smooth optimization 1

one-shot optimization 398
optimal control 76, 113
optimal sparse control 169
ordinary differential systems 74

*p*-Laplacian 138
penalty method 399
phase field 368
pointwise control 34
population dynamics 91
porous media 238
pseudo semi-smooth Newton 16

reaction–diffusion systems 95

semi-smooth Newton 174
semilinear parabolic equation 170
space-time variational formulation 288
stabilization 368
stationarity conditions 17
Steiner tree 197
Stokes system 238
strange term 243
super-solutions 101

traveling waves 96

uncertainty quantification 420

vanishing noise 399
variational discretization 116

wave equation 33
Wolbachia 74

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

# Radon Series on Computational and Applied Mathematics

**Volume 28**
*Multiphysics Phase-Field Fracture. Modeling, Adaptive Discretizations, and Solvers*
Thomas Wick, 2020
ISBN: 978-3-11-049656-7, e-ISBN: 978-3-11-049656-7

**Volume 27**
*Multivariate Algorithms and Information-Based Complexity*
Fred J. Hickernell, Peter Kritzer (Eds.), 2020
ISBN: 978-3-11-063311-5, e-ISBN: 978-3-11-063546-1

**Volume 26**
*Discrepancy Theory*
Dmitriy Bilyk, Josef Dick, Friedrich Pillichshammer (Eds.), 2020
ISBN: 978-3-11-065115-7, e-ISBN: 978-3-11-065258-1

**Volume 25**
*Space-Time Methods. Applications to Partial Differential Equations*
Ulrich Langer, Olaf Steinbach (Eds.), 2019
ISBN: 978-3-11-054787-0, e-ISBN: 978-3-11-054848-8

**Volume 24**
*Maxwell's Equations. Analysis and Numerics*
Ulrich Langer, Dirk Pauly, Sergey I. Repin (Eds.), 2019
ISBN: 978-3-11-054264-6, e-ISBN: 978-3-11-054361-2

**Volume 23**
*Combinatorics and Finite Fields. Difference Sets, Polynomials, Pseudorandomness and Applications*
Kai-Uwe Schmidt, Arne Winterhof (Eds.), 2019
ISBN: 978-3-11-064179-0, e-ISBN: 978-3-11-064209-4

**Volume 22**
*The Radon Transform. The First 100 Years and Beyond*
Ronny Ramlau, Otmar Scherzer (Eds.), 2019
ISBN: 978-3-11-055941-5, e-ISBN: 978-3-11-056085-5

www.degruyter.com

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 3:19 PM via . All use subject to https://www.ebsco.com/terms-of-use