International Series of Numerical Mathematics

164

Kristian Bredies
Christian Clason
Karl Kunisch
Gregory von Winckel
Editors

Control and Optimization with PDE Constraints

Birkhäuser

Birkhäuser

ISNM

International Series of Numerical Mathematics

Volume 164

For further volumes:
www.springer.com/series/4819

Kristian Bredies • Christian Clason •
Karl Kunisch • Gregory von Winckel
Editors

Control and Optimization with PDE Constraints

Birkhäuser

Editors

Kristian Bredies
Institute for Mathematics and Scientific Computing
University of Graz
Graz, Austria

Karl Kunisch
Institute for Mathematics and Scientific Computing
University of Graz
Graz, Austria

Christian Clason
Institute for Mathematics and Scientific Computing
University of Graz
Graz, Austria

Gregory von Winckel
Institute for Mathematics and Scientific Computing
University of Graz
Graz, Austria

ISBN 978-3-0348-0630-5 ISBN 978-3-0348-0631-2 (eBook)
DOI 10.1007/978-3-0348-0631-2
Springer Basel Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013940252

Mathematics Subject Classification (2010): Primary 49-XX, 35-XX, Secondary 35J25, 35K20, 35R60, 49J20, 49K20, 49M05, 49M15, 49M25, 65K10, 65M60, 65N15, 65N30

Printed on acid-free paper

Springer Basel is part of Springer Science+Business Media (www.birkhauser-science.com)

Preface

The articles contained in this volume had their genesis in presentations given at the *International Workshop on Control and Optimization of PDEs*, held at the Bildungshaus Mariatrost from October 10 to 14, 2011. These contributions, made by internationally well-known researchers in applied mathematics, cover a wide variety of topics in PDE-constrained optimization and control such as multiscale and stochastic problems, model reduction and domain decomposition, control of wave equations, delay systems, and nonsmooth problems. The applications considered range from control of quantum systems over diffusion in porous media to calibration of option pricing models.

Control and optimization of PDEs is an active field, with research increasingly going beyond standard settings and approaches. Let us mention just a few examples. Model reduction, which has been very successful for linear elliptic and parabolic problems, is now extended to hyperbolic and to nonlinear problems. Similarly, techniques from control of dynamical systems and nonlinear optimization are being applied to optimal control problems for PDEs. Another area of increasing activity is concerned with optimal control problems in Banach or even metric spaces (such as the space L^p for $0 \leq p < 1$) rather than the standard Hilbert spaces, leading to the development of novel approaches and opening new applications such as sparsity constraints.

The workshop was successful in bringing together researchers working at the forefront of their respective fields, from theoretical analysis to numerical realization and applications, which frequently do not have strong interactions. Consequently, this book addresses researchers in all areas of control and optimization of systems governed by differential equations.

The editors express their gratitude to the contributors of this volume as well as to Birkhäuser/Springer Basel for publishing it. Financial support of the European Science Foundation within the framework "Optimization with PDE constraints", the Austrian Science Fund (FWF) under grant F32 (SFB "Mathematical Optimization and Applications in Biomedical Sciences"), Karl-Franzens-University of Graz and the government of Styria is gratefully acknowledged.

Graz, Austria C. Clason
Graz, Austria K. Kunisch

Contents

Contributors

Alessandro Alla Università degli studi di Roma “La Sapienza”, Roma, Italy

H.T. Banks Center for Research in Scientific Computation, Center for Quantitative Science in Biomedicine, North Carolina State University, Raleigh, NC, USA

Susanne Beckers Faculty of Mathematics, Lothar Collatz School for Computing in Science, Universität Hamburg, Hamburg, Germany

Maurizio Falcone Università degli studi di Roma “La Sapienza”, Roma, Italy

Kazufumi Ito Department of Mathematics and Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC, USA

Yvon Maday UMR 7598, Laboratoire Jacques-Louis Lions, UPMC Univ. Paris 06, Paris, France; Division of Applied Mathematics, Brown University, Providence, RI, USA

Zhiping Rao Commands (ENSTA ParisTech, INRIA Saclay), Palaiseau Cedex, France

Mohamed-Kamel Riahi UMR 7598, Laboratoire Jacques-Louis Lions, UPMC Univ. Paris 06, Paris, France

Danielle Robbins Center for Research in Scientific Computation, Center for Quantitative Science in Biomedicine, North Carolina State University, Raleigh, NC, USA

Ekkehard W. Sachs Fachbereich IV, Abteilung Mathematik, Universität Trier, Trier, Germany

Julien Salomon CEREMADE, Université Paris-Dauphine, Paris, France

Matthias Schu Fachbereich IV, Abteilung Mathematik, Universität Trier, Trier, Germany

Alina Studinger Department of Mathematics and Statistics, University of Constance, Konstanz, Germany

Karyn L. Sutton Center for Research in Scientific Computation, Center for Quantitative Science in Biomedicine, North Carolina State University, Raleigh, NC, USA

Gabriel Turinici CEREMADE, Université Paris Dauphine, Paris, France

Stefan Volkwein Department of Mathematics and Statistics, University of Constance, Konstanz, Germany

Gregory von Winckel Institut für Mathematik und Wissenschaftliches Rechnen, Karl-Franzens-Universität Graz, Graz, Austria

Winnifried Wollner Department of Mathematics, University of Hamburg, Hamburg, Germany

Masahiro Yamamoto Graduate School of Mathematical Sciences, The University of Tokyo, Tokyo, Japan

Hasnaa Zidani Commands (ENSTA ParisTech, INRIA Saclay), Palaiseau Cedex, France

An Adaptive POD Approximation Method for the Control of Advection-Diffusion Equations

Alessandro Alla and Maurizio Falcone

Abstract We present an algorithm for the approximation of a finite horizon optimal control problem for advection-diffusion equations. The method is based on the coupling between an adaptive POD representation of the solution and a Dynamic Programming approximation scheme for the corresponding evolutive Hamilton–Jacobi equation. We discuss several features regarding the adaptivity of the method, the role of error estimate indicators to choose a time subdivision of the problem and the computation of the basis functions. Some test problems are presented to illustrate the method.

Keywords Optimal Control · Proper orthogonal decomposition · Hamilton–Jacobi equations · Advection-diffusion equations

Mathematics Subject Classification (2010) Primary 49J20 · 49L20 · Secondary 49M25

1 Introduction

The approximation of optimal control problems for evolutionary partial differential equations of parabolic and hyperbolic type is a very challenging topic with a strong impact on industrial applications. Although there is a large number of papers dealing with several aspects of control problems from controllability to optimal control, the literature dealing with the numerical approximation of such huge problems is rather limited. It is worth to note that when dealing with optimal control problems for

The authors wish to acknowledge the support obtained by the following grants: ESF-OPTPDE Network, ITN—Marie Curie Grant n. 264735-SADCO and PRIN 2009 "Metodi Innovativi per il Calcolo Scientifico".
The authors also wish to thank the CASPUR Consortium for its technical support.

A. Alla · M. Falcone (✉)
Università degli studi di Roma "La Sapienza", Piazzale Aldo Moro, 2, 0010 Roma, Italy
e-mail: falcone@mat.uniroma1.it

A. Alla
e-mail: alla@mat.uniroma1.it

K. Bredies et al. (eds.), *Control and Optimization with PDE Constraints*,
International Series of Numerical Mathematics 164,
DOI 10.1007/978-3-0348-0631-2_1, © Springer Basel 2013

parabolic equations we can exploit the regularity of the solutions, regularity which is lacking for many hyperbolic equations. We also recall that the main tools is still given by the Pontryagin maximum principle (see e.g. [14]). This is mainly due to the fact that the discretization of partial differential equations already involves a large number of variables so that the resulting finite dimensional optimization problem easily reaches the limits of what one can really compute. The forward-backward system which describes Pontryagin's optimality condition is certainly below that limit. However just solving that system one is using necessary conditions for optimality so, in principle, there is no guarantee that these are optimal controls. By this approach for general nonlinear control problems we can obtain just open-loop control. One notable exception is the linear quadratic regulator problem for which we have a closed-loop solution given by the Riccati equation. This explains why the most popular example for the control of evolutive partial differential equations is the control of the heat equation subject to a quadratic cost functional.

In recent years, new tools have been developed to deal with optimal control problems in infinite dimension. In particular, new techniques emerged to reduce the number of dimensions in the description of the dynamical system or, more in general, of the solution of the problem that one is trying to optimize. These methods are generally called *reduced-order methods* and include for example the POD (Proper Orthogonal Decomposition) method and reduced basis approximation (see [12]). The general idea for all this method is that, when the solution are sufficiently regular, one can represent them via Galerkin expansion so that the number of variables involved in this discretization will be strongly reduced. In some particular case, as for the heat equation, even 5 basis functions will suffice to have a rather accurate POD representation of the solution. Having this in mind, it is reasonable to start thinking to a different approach based on Dynamic Programming (DP) and Hamilton–Jacobi-Bellman equations (HJB). In this new approach we will first develop a reduced basis representation of the solution along a reference trajectory and then use this basis to set-up a control problem in the new space of coordinates. The corresponding Hamilton–Jacobi equation will just need 3–5 variables to represent the state of the system. Moreover, by this method one can obtain optimal control in feedback form looking at the gradient of the value function.

However, the solution of HJB equation is not an easy task from the numerical point of view: the analytical solution of the HJB equation is non regular (typically, just Lipschitz continuous). Optimal control problems for ODEs were solved by Dynamic Programming, both analytically and numerically (see [1] for a general presentation of this theory). From the numerical point of view, this approach has been developed for many classical control problems obtaining convergence results and a-priori error estimates ([4, 6] and the book [5]). Although this approach suffers from the curse-of-dimensionality, some algorithms in high-dimension are now available ([3] and [2]), and the coupling with POD representation techniques will allow to attack by this technique optimal control problems in infinite dimension.

To set this paper into perspective we must say that a first tentative step in this direction has been made by Kunisch and co-authors in a series of papers [7, 8] for diffusion dominated equations. In particular, in the paper by Kunisch, Volkwein and

Xie [10, 11] one can see a feedback control approach based on coupling between POD basis approximation and HJB equations for the viscous Burgers' equation. Our contribution here is twofold. The first novelty is that we deal with advection-diffusion equations. The solutions to these equations exhibit low regularity properties with respect to non degenerate diffusion equations so that a rather large number of POD basis functions will be required to obtain a good approximation if we want to compute the POD basis just once. Naturally, this increases the number of variables in the HJB approach and constitutes a real bottle-neck. In order to apply the Dynamic Programming approach to this problem we have developed an adaptive technique which allows to recompute the POD basis on different sub-intervals in order to have always accurate results without an increase of the number of basis functions. The second contribution of this paper is the way the sub-intervals are determined. In fact, we do not use a simple uniform subdivision but rather decide to recompute the POD basis when an error indicator (detailed in Sect. 4) is beyond a given threshold. As we will show in the sequel, this procedure seems to be rather efficient and accurate to deal with these large scale problems.

2 The POD Approximation Method for Evolutive PDEs

We briefly describe some important features of the POD approximation, more details as well as precise results can be found in the notes by Volkwein [15]. Let us consider a matrix $Y \in \mathbb{R}^{m\times n}$, with rank $d \leq \min\{m,n\}$. We will call y_j the jth column of the matrix Y. We are looking for an orthonormal basis $\{\psi_i\}_{i=1}^{\ell} \in \mathbb{R}^m$ with $\ell \leq n$ such that the minimum of the following functional is reached:

$$J(\psi_1,\ldots,\psi_\ell) = \sum_{j=1}^{n} \left\| y_j - \sum_{i=1}^{\ell} \langle y_j, \psi_i\rangle \psi_i \right\|^2 . \tag{2.1}$$

The solution of this minimization problem is given in the following theorem

Theorem 1 *Let $Y = [y_1,\ldots,y_n] \in \mathbb{R}^{m\times n}$ be a given matrix with rank $d \leq \min\{m,n\}$. Further, let $Y = \Psi\Sigma V^T$ be the Singular Value Decomposition (SVD) of Y, where $\Psi = [\psi_1,\ldots,\psi_m] \in \mathbb{R}^{m\times m}$, $V = [v_1,\ldots,v_n] \in \mathbb{R}^{n\times n}$ are orthogonal matrices and the matrix $\Sigma \in \mathbb{R}^{m\times n}$ is diagonal, $\Sigma = \mathrm{diag}\{\sigma_1,\ldots,\sigma_m\}$. Then, for any $\ell \in \{1,\ldots,d\}$ the solution to (2.1) is given by the left singular vectors $\{\psi_i\}_{i=1}^{\ell}$, i.e, by the first ℓ columns of Ψ.*

We will call the vectors $\{\psi_i\}_{i=1}^{\ell}$ *POD basis* of rank ℓ.

This idea is really useful, in fact we get a solution solving an equation whose dimension is decreased with respect to the initial one. Whenever it's possible to compute a POD basis of rank ℓ, we get a problem with much smaller dimension of the starting one due to the fact ℓ is properly chosen very small.

Let us consider the following ODEs system

$$\begin{cases} \dot{y}(s) = Ay(s) + f\big(s, y(s)\big), \quad s \in (0, T], \\ y(0) = y_0, \end{cases} \tag{2.2}$$

where $y_0 \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times m}$ and $f : [0, T] \times \mathbb{R}^m \to \mathbb{R}^m$ is continuous and locally Lipschitz to ensure uniqueness.

The system (2.2) can also be interpreted as a semidiscrete problem, where the matrix A represents the discretization in space of an elliptic operator, for instance the Laplacian. To compute the POD basis functions, first of all we have to construct a time grid $0 \le t_1 \le \cdots \le t_n = T$ and we suppose to know the solution of (2.2) at given time t_j, $j = 1, \dots, N$. We call *snapshots* the solution at those fixed times. For the moment we will not deal with the problem of selecting the snapshots sequence which is a difficult problem in itself; we refer the interested readers to [9]. As soon as we get the snapshots sequence, by Theorem 1, we will be able to compute our POD basis, namely, $\{\psi_j\}_{j=1}^{\ell}$.

Let us suppose we can write the solution in reduced form as

$$y^\ell(s) = \sum_{j=1}^{\ell} y_j^\ell(s)\psi_j = \sum_{j=1}^{\ell} \big\langle y^\ell(s), \psi_j\big\rangle \psi_j, \quad \forall s \in [0, T].$$

Substituting this formula into (2.2) we obtain the reduced dynamics

$$\begin{cases} \sum\limits_{j=1}^{\ell} \dot{y}_j^\ell(s)\psi_j = \sum\limits_{j=1}^{\ell} y_j^\ell(s) A\psi_j + f\big(s, y^\ell(s)\big), \quad s \in (0, T], \\ \sum\limits_{j=1}^{\ell} y_j^\ell(0)\psi_j = y_0. \end{cases} \tag{2.3}$$

We note that our new problem (2.3) is a problem for the $\ell \le m$ coefficient functions $y_j^\ell(s)$, $j = 1, \dots, \ell$. Thus, the problem is low dimensional and with compact notation we get:

$$\begin{cases} \dot{y}^\ell(s) = A^\ell y^\ell(s) + F\big(s, y^\ell(s)\big), \\ y^\ell(0) = y_0^\ell, \end{cases}$$

where

$$A^\ell \in \mathbb{R}^{\ell\times\ell} \quad \text{with } \big(A^\ell\big)_{ij} = \langle A\psi_i, \psi_j\rangle,$$

$$y^\ell = \begin{pmatrix} y_1^\ell \\ \vdots \\ y_\ell^\ell \end{pmatrix} : [0, T] \to \mathbb{R}^\ell,$$

$$F = (F_1, \dots, F_\ell)^T : [0, T] \times \mathbb{R}^\ell \to \mathbb{R}^\ell,$$

$$F_i(s,y) = \left\langle f\left(s, \sum_{j=1}^{\ell} y_j \psi_j\right), \psi_i \right\rangle \quad \text{for } s \in [0,T],\ y = (y_1, \dots y_\ell) \in \mathbb{R}^\ell.$$

Finally, we obtain the representation of y_0 in $\mathbb{R}^\ell$

$$y_0^\ell = \begin{pmatrix} \langle y_0, \psi_1 \rangle \\ \vdots \\ \langle y_0, \psi_\ell \rangle \end{pmatrix} \in \mathbb{R}^\ell.$$

In order to apply the POD method to our optimal control problem, the number ℓ of POD basis functions is crucial. In particular we would like to keep ℓ as low as possible still capturing the behavior of the original dynamics. The problem is to define an indicator of the accuracy of our POD approximation. A good choice for this indicator is the following ratio

$$\mathcal{E}(\ell) = \frac{\sum_{i=1}^{\ell} \sigma_i}{\sum_{i=1}^{d} \sigma_i}, \tag{2.4}$$

where the σ_i are the singular value obtained by the SVD.

As much $\mathcal{E}(\ell)$ is close to one as much our approximation will be improved. This is strictly related to the truncation error due to the projection of y_j onto the space generated by the orthonormal basis $\{\psi\}_{i=1}^{\ell}$, in fact:

$$J(\psi_1, \dots, \psi_\ell) = \sum_{j=1}^{n} \left\| y_j - \sum_{i=1}^{\ell} \langle y_j, \psi_i \rangle \psi_i \right\|^2 = \sum_{i=\ell+1}^{d} \sigma_i^2.$$

3 An Optimal Control Problem

We will present this approach for the finite horizon control problem. Consider the controlled system

$$\begin{cases} \dot{y}(s) = f\big(y(s), u(s), s\big), & s \in (t,T], \\ y(t) = x \in \mathbb{R}^n, \end{cases} \tag{3.1}$$

we will denote by $y : [t,T] \to \mathbb{R}^n$ its solution, by u the control $u : [t,T] \to \mathbb{R}^m$, $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, $s \in (t,T]$ and by

$$\mathcal{U} = \big\{u : [0,T] \to U\big\}$$

the set of admissible controls where $U \subset \mathbb{R}^m$ is a compact set. Whenever we want to emphasize the dependence of the solution from the control u we will write $y(t;u)$. Assume that there exists a unique solution trajectory for (3.1) provided the controls are measurable (a precise statement can be found in [1]). For the finite horizon optimal control problem the cost functional will be given by

$$\min_{u\in\mathcal{U}} J_{x,t}(u) := \int_t^T L\big(y(s,u),u(s),s\big)e^{-\lambda s}\,ds + g\big(y(T)\big) \tag{3.2}$$

where $L:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$ is the running cost and $\lambda\geq 0$ is the discount factor.

The goal is to find a state-feedback control law $u(t)=\Phi(y(t),t)$, in terms of the state equation $y(t)$, where Φ is the feedback map. To derive optimality conditions we use the well-known *dynamic programming principle* due to Bellman (see [1]). We first define the value function

$$v(x,t) := \inf_{u\in\mathcal{U}} J_{x,t}(u). \tag{3.3}$$

Proposition 3.1 (DPP) *For all* $x\in\mathbb{R}^n$ *and* $0\leq\tau\leq t$, *we have*

$$v(x,t) = \min_{u\in\mathcal{U}}\left\{\int_t^\tau L\big(y(s),u(s),s\big)e^{-\lambda s}\,ds + v(y,t-\tau)\right\}. \tag{3.4}$$

Due to (3.4) we can derive the *Hamilton–Jacobi-Bellman* equations (HJB):

$$-\frac{\partial v}{\partial t}(y,t) = \min_{u\in U}\big\{L(y,u,t)+\nabla v(y,t)\cdot f(y,u,t)\big\}. \tag{3.5}$$

This is nonlinear partial differential equation of the first order which is hard to solve analytically although a general theory of weak solutions is available [1]. Rather we can solve it numerically by means of a finite differences or semi-Lagrangian schemes (see the book [5] for a comprehensive analysis of approximation schemes for Hamilton–Jacobi equations). For a semi-Lagrangian discretization one starts by a discrete version of (HJB) by discretizing the underlined control problem and then project the semi-discrete scheme on a grid obtaining the fully discrete scheme

$$\begin{cases} v_i^{n+1} = \min\limits_{u\in U}\big[\Delta t\, L(x_i,n\Delta t,u) + I\big[v^n\big]\big(x_i+\Delta t\, F(x_i,t_n,u)\big)\big],\\ v_i^0 = g(x_i)\end{cases}$$

with $x_i=i\Delta x$, $t_n=n\Delta t$, $v_i^n := v(x_i,t_n)$ and $I[\cdot]$ is an interpolation operator which is necessary to compute the value of v^n at the point $x_i+\Delta t\, F(x_i,t_n,u)$ (in general, this point will not be a node of the grid). The interested reader will find in [6] a detailed presentation of the scheme and a priori error estimates for its numerical approximation.

Note that, we also need to compute the minimum in order to get the value v_i^{n+1}. Since v^n is not a smooth function, we compute the minimum by means of a minimization method which does not use derivatives (this can be done by the Brent algorithm as in [3]).

As we already noted, the HJB allows to compute the optimal feedback via the value function, but there are two major difficulties: the solution of an HJB equation are in general non-smooth and the approximation in high dimension is not feasible. The request to solve an HJB in high dimension comes up naturally whenever we want to control evolutive PDEs. Just to give an idea, if we build a grid in [0, 1] × [0, 1] with a discrete step $\Delta x = 0.01$ we have 10^4 nodes: to solve an HJB in that dimension is simply impossible. Fortunately, the POD method allows us to obtain reduced models even for complex dynamics. Let us focus on the following abstract problem:

$$\begin{cases} \dfrac{d}{ds}\big\langle y(s), \varphi\big\rangle_H + a\big(y(s), \varphi\big) = \big\langle B\big(u(s), \varphi\big)\big\rangle_{V',V} \quad \forall \varphi \in V, \\ y(t) = y_0 \in H, \end{cases} \tag{3.6}$$

where $B : U \to V'$ is a linear and continuous operator. We assume that a space of admissible controls $\mathcal{U}_{ad}$ is given in such a way that for each $u \in \mathcal{U}_{ad}$ and $y_0 \in H$ there exists a unique solution y of (3.6). V and H are two Hilbert spaces, with $\langle \cdot, \cdot \rangle_H$ we denote the scalar product in H; $a : V \times V \to \mathbb{R}$: is symmetric coercive and bilinear. Then, we introduce the cost functional of the finite horizon problem

$$\mathcal{J}_{y_0,t}(u) := \int_t^T L\big(y(s), u(s), s\big) e^{-\lambda s}\, ds + g\big(y(T)\big),$$

where $L : V \times U \times [0, T] \to \mathbb{R}$. The optimal control problem is

$$\begin{aligned} &\min_{u \in \mathcal{U}_{ad}} \mathcal{J}_{y_0,t}(u) \\ &\text{subject to} \quad y \in W_{loc}(0, T; V) \times \mathcal{U} \quad \text{solves (3.6)} \end{aligned} \tag{3.7}$$

with $W_{loc}(0, T) = \bigcap_{T>0} W(0, T)$, where $W(0, T)$ is the standard Sobolev space

$$W(0, T) = \big\{\varphi \in L^2(0, T; V), \varphi_t \in L^2\big(0, T; V'\big)\big\}.$$

The model reduction approach for an optimal control problem (3.7) is based on the Galerkin approximation of dynamic with some information on the controlled dynamic (snapshots). To compute a POD solution for (3.7) we make the following ansatz

$$y^\ell(x, s) = \sum_{i=1}^{\ell} w_i(s) \psi_i(x), \tag{3.8}$$

where $\{\psi\}_{i=1}^{\ell}$ is the POD basis computed as in the previous section.

We introduce mass and stiffness matrices:

$$M = (m_{ij}) \in \mathbb{R}^{\ell \times \ell} \quad \text{with } m_{ij} = \langle \psi_j, \psi_i \rangle_H,$$

$$S = (s_{ij}) \in \mathbb{R}^{\ell \times \ell} \quad \text{with } m_{ij} = a(\psi_j, \psi_i),$$

and the control map $b : U \to \mathbb{R}^\ell$ is defined by:

$$u \to b(u) = \big(b(u)_i\big) \in \mathbb{R}^\ell \quad \text{with } b(u)_i = \langle Bu, \psi_i \rangle_H.$$

The coefficients of the initial condition $y^\ell(0) \in \mathbb{R}^\ell$ are determined by $w_i(0) = (w_0)_i = \langle y_0, \psi \rangle_X$, $1 \leq i \leq \ell$, and the solution of the reduced dynamic problem is denoted by $w^\ell(s) \in \mathbb{R}^\ell$. Then, the Galerkin approximation is given by

$$\min J^\ell_{w_0^\ell,t}(u) \tag{3.9}$$

with $u \in \mathcal{U}_{ad}$ and w solving

$$\begin{cases} \dot{w}^\ell(s) = F\big(w^\ell(s), u(s), s\big), & s > 0, \\ w^\ell(0) = w_0^\ell. \end{cases} \tag{3.10}$$

The cost functional is defined as

$$J^\ell_{w_0^\ell,t}(u) = \int_0^T L\big(w^\ell(s), u(s), s\big)e^{-\lambda s}\, dt + g\big(w^\ell(T)\big),$$

with w^ℓ and y^ℓ linked to (3.8) and the nonlinear map $F : \mathbb{R}^\ell \times U \to \mathbb{R}^\ell$ is given by

$$F\big(w^\ell, u, s\big) = M^{-1}\big(-Sw^\ell(s) + b\big(u(s)\big)\big).$$

The value function v^ℓ, defined for the initial state $w_0 \in \mathbb{R}^\ell$, is

$$v^\ell\big(w_0^\ell, t\big) = \inf_{u \in \mathcal{U}_{ad}} J^\ell_{w_0^\ell,t}(u)$$

and w^ℓ solves (3.9) with the control u and initial condition w_0.

We give an idea how we have computed the intervals for reduced HJB. HJBs are defined in $\mathbb{R}^n$, but we have restricted our numerical domain Υ_h which is a bounded subset of $\mathbb{R}^n$. This is justified since $y + \Delta t F(y, u) \in \Upsilon_h$ for each $y \in \Upsilon_h$ and $u \in \mathcal{U}_{ad}$. We can chose $\Upsilon_h = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_\ell, b_\ell]$ with $a_1 \geq a_2 \geq \cdots \geq a_\ell$. How should we compute these intervals $[a_i, b_i]$?

Ideally the intervals should be chosen so that the dynamics contains all the components of the controlled trajectory. Moreover, they should be encapsulated because we expect that their importance should decrease monotonically with their index and that our interval lengths decrease quickly.

Let us suppose to discretize the space control $U = \{u_1, \ldots, u_M\}$ where U is symmetric, to be more precise if $\bar{u} \in U \Rightarrow -\bar{u} \in U$. Hence, if

$$y^\ell(s) = \sum_{i=1}^{\ell} \big\langle y(s), \psi_i \big\rangle \psi_i = \sum_{i=1}^{\ell} w_i(s)\psi_i,$$

as a consequence, the coefficients $w_i(s) \in [a_i, b_i]$. We consider the trajectories solution $y(s, u_j)$ such that the control is constant $u(s) \equiv u_j$ for each t_j, $j = 1, \ldots, M$. Then, we have

$$y^\ell(s, u_j) = \sum_{i=1}^{\ell} \big\langle y(s, u_j), \psi_i \big\rangle \psi_i.$$

We write $y^\ell(s,u_j)$ to stress the dependence on the constant control u_j. Each trajectory $y^\ell(s,u_j)$ has some coefficients $w_i^{(j)}(t)$ for $i=1,\dots,\ell$, $j=1,\dots,M$. The coefficients $w_i^{(j)}(s)$ will belong to intervals of the type $[\underline{\mathrm{w}}_i^{(j)}, \overline{\mathrm{w}}_i^{(j)}]$ where we chose for $i=1,\dots,\ell$, a_i, b_i such that:

$$\begin{aligned} a_i &\equiv \min\left\{\underline{\mathrm{w}}_i^{(1)},\dots,\underline{\mathrm{w}}_i^{(M)}\right\},\\ b_i &\equiv \max\left\{\overline{\mathrm{w}}_i^{(1)},\dots,\overline{\mathrm{w}}_i^{(M)}\right\}. \end{aligned}$$

Then, we have a method to compute the intervals and we turn our attention to the numerical solution of an optimal control problem for the evolutive equation, as we will see in the following section.

4 Adapting POD Approximation

We now present an adaptive method to compute a POD basis. Since our final goal is to obtain the optimal feedback law by means of HJB equations, we will have a big constraint on the number of variables in the state space for numerical solution of an HJB.

We will see that, for a parabolic equation, one can try to solve the problem with only three/four POD basis functions; they are enough to describe the solution in a rather accurate way. In fact the singular values decay pretty soon and it's easier to work with a really low-rank dimensional problem.

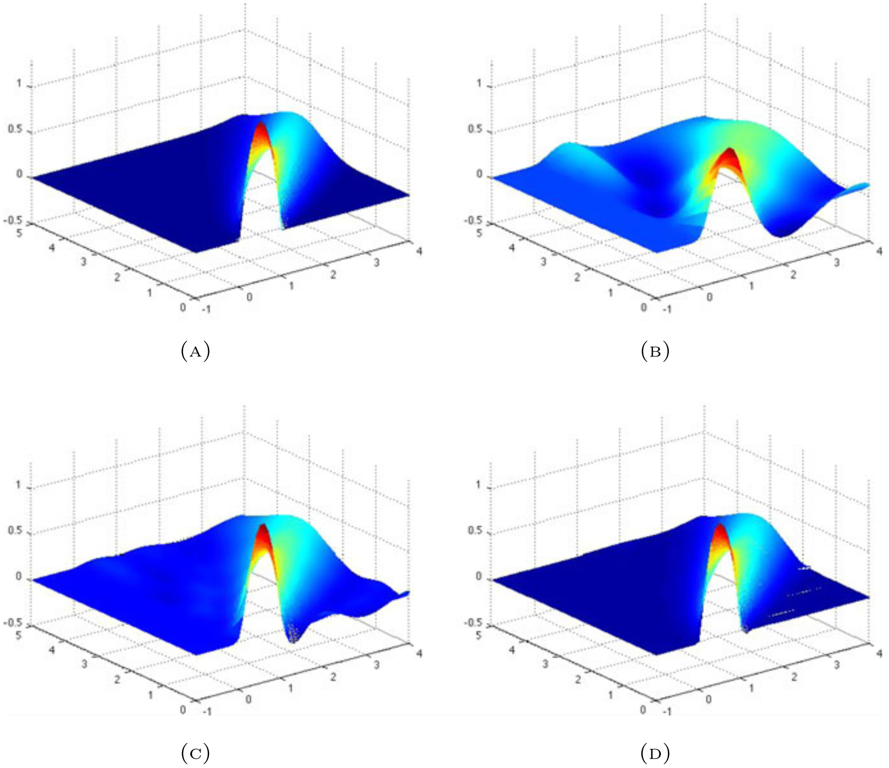
On the contrary, hyperbolic equations do not have this nice property for their singular values and they will require a rather large set of POD basis functions to get accurate results. Note that we can not follow the approach suggested in [13] because we can not add more basis functions when it turns to be necessary due to the constraint already mentioned. Then, it is quite natural to split the problem into subproblems having different POD basis functions. The crucial point is to decide the splitting in order to have the same number of basis functions in each subdomain with a guaranteed accuracy in the approximation.

Let us first give an illustrative example for the parabolic case, considering a 1D advection-diffusion equation:

$$\begin{cases} y_s(x,s)-\varepsilon y_{xx}(x,s)+c y_x(x,s)=0,\\ y(x,0)=y_0(x), \end{cases} \tag{4.1}$$

with $x\in[a,b]$, $s\in[0,T]$, $\varepsilon, c\in\mathbb{R}$.

We use a finite difference approximation for this equation based on an explicit Euler method in time combined with the standard centered approximation of the second order term and with an up-wind correction for the advection term. The snapshots will be taken from the sequence generated by the finite difference method. The final



(A) (B)

(C) (D)

Fig. 1 Equation (4.1): (**a**) Solved with finite difference; (**b**) POD-Galerkin approximation with 3 POD basis; (**c**) Solved via POD-Galerkin approximation with 5 POD basis; (**d**) Adapting 3 POD basis functions

time is $T = 5$, moreover $a = -1$, $b = 4$. The initial condition is $y_0(x) = 5x - 5x^2$, when $0 \leq x \leq 1$, 0 otherwise.

For $\varepsilon = 0.05$ and $c = 1$ with only 3 POD basis functions, the approximation fails (see Fig. 1). Note that in this case the advection is dominating the diffusion, a low number of POD basis functions will not suffice to get an accurate approximation (Fig. 1b). However, the adaptive method which only uses 3 POD basis functions will give accurate results (Fig. 1d).

The idea which is behind the adaptive method is the following: we do not consider all the snapshots together in the whole interval $[0, T]$ but we group them. Instead of taking into account the whole interval $[0, T]$, we prefer to split it in sub-intervals

$$[0, T] = \bigcup_{k=0}^{K} [T_k, T_{k+1}],$$

where K is a-priori unknown, $T_0 = 0$, $T_K = T$ and $T_k = t_i$ for some i. In this way, choosing properly the length of the kth interval $[T_k, T_{k+1}]$, we consider only the snapshots falling in that sub-interval, typically there will be at least three snapshots in every sub-interval. Then we have enough information in every sub-interval and we can apply the standard routines (explained in Sect. 2) to get a "local" POD basis.

Now let us explain how to divide our time interval $[0,T]$. We will choose a parameter to check the accuracy of the POD approximation and define a threshold. Above that threshold we loose in accuracy and we need to compute a new POD basis. A good parameter to check the accuracy is $\mathcal{E}(\ell)$ (see (2.4)), as it was suggested by several authors. The method to define the splitting of $[0,T]$ and the size of every sub-interval works as follows. We start computing the SVD of the matrix Y that gives us information about our dynamics in the whole time interval. We check the accuracy at every t_i, $i=1,\dots N$, and if at t_k the indicator is above the tolerance we set $T_1 = t_k$ and we divide the interval in two parts, $[0,T_1)$ and $(T_1,T]$. Now we just consider the snapshots related the solution up to the time T_1. Then we iterate this idea until the indicator is below the threshold. When the first interval is found, we restart the procedure in the interval $[T_1,T]$ and we stop when we reach the final time T. Note that the extrema of every interval coincide by construction with one of our discrete times $t_i = i\Delta t$ so that the global solution is easily obtained linking all the sub-problems which always have a snapshot as initial condition. A low value for the threshold will also guarantee that we will not have big jumps passing from one sub-interval to the next.

This idea can be applied also when we have a controlled dynamic (see (5.1)). First of all we have to decide how to collect the snapshots, since the control $u(t)$ is completely unknown. One can make a guess and use the dynamics and the functional corresponding to that guess, by these information we can compute the POD basis. Once the POD basis is obtained we will get the optimal feedback law after having solved a reduced HJB equation as we already explained. Let us summarize the method in the following step-by-step presentation.

Algorithm

```
Start: Initialization
Step 1: collect the snapshots in $[0,T]$
Step 2: divide $[0,T]$ according to $\mathcal{E}(\ell)$
For i=0 to N-1
Do
  Step 3: apply SVD to get the POD basis in each
      sub-interval $[t_i,t_{i+1}]$
  Step 4: discretize the space of controls
  Step 5: project the dynamics onto the (reduced) POD space
  Step 6: select the intervals for the POD reduced
      variables
  Step 7: solve the corresponding HJB in the reduced space
      for the interval $[t_i,t_{i+1}]$
  Step 8: go back to the original coordinate space
End
```

5 Numerical Experiments

In this section we present some numerical tests for the controlled heat equation and for the advection-diffusion equation with a quadratic cost functional. Consider the following advection-diffusion equation:

$$\begin{cases} y_s(x,s) - \varepsilon y_{xx}(x,s) + c y_x(x,s) = u(s), \\ y(x,0) = y_0(x), \end{cases} \tag{5.1}$$

with $x \in [a,b]$, $s \in [0,T]$, $\varepsilon \in \mathbb{R}_+$ and $c \in \mathbb{R}$. Note that changing the parameters c and ε we can obtain the heat equation ($c = 0$) and the advection equation ($\varepsilon = 0$). The functional to be minimized is

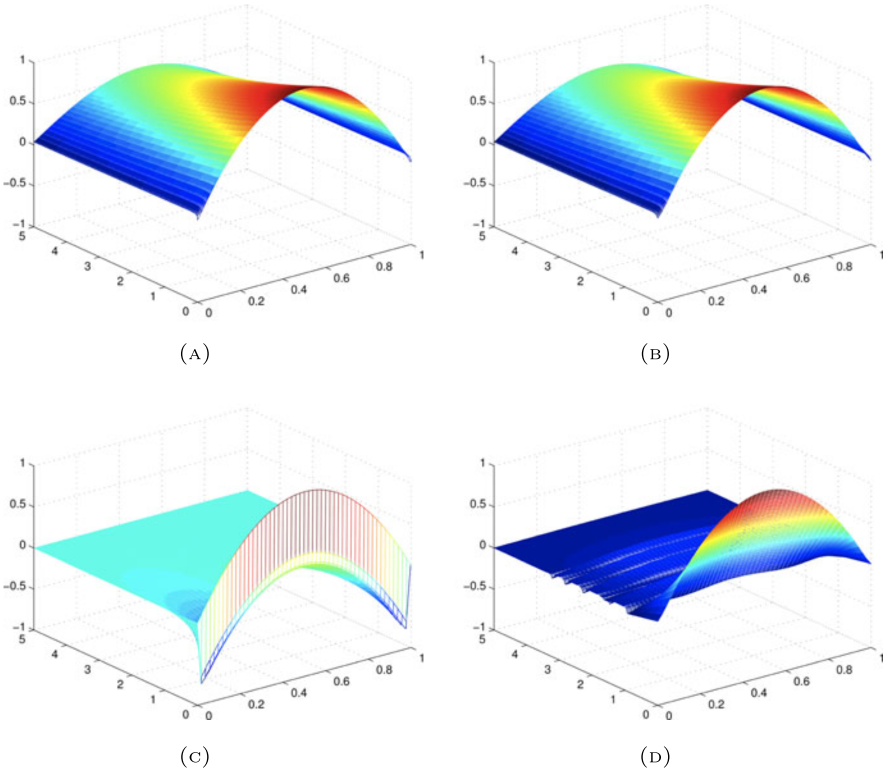
$$J_{y_0,t}\big(u(\cdot)\big) = \int_0^T \big\| y(x,s) - \hat{y}(x,s) \big\|^2 + R \big\| u(s) \big\|^2 ds, \tag{5.2}$$

i.e., we want to stay close to a reference trajectory $\hat{y}$ while minimizing the norm of u. Note that we dropped the discount factor setting $\lambda = 0$. Typically in our test problems $\hat{y}$ is obtained by applying a particular control $\hat{u}$ to the dynamics. The numerical simulations reported in this paper have been made on a server SUPERMICRO 8045C-3RB with 2 cpu Intel Xeon Quad-Core 2.4 Ghz and 32 GB RAM under SLURM (https://computing.llnl.gov/linux/slurm/).

Test 1: Heat Equation with Smooth Initial Data We compute the snapshots with a centered/forward Euler scheme with space step $\Delta x = 0.02$, and time step $\Delta t = 0.012$, $\varepsilon = 1/60$, $c = 0$, $R = 0.01$ and $T = 5$. The initial condition is $y_0(x) = 5x - 5x^2$, and $\hat{y}(x,s) = 0$. In Fig. 2 we compare four different approximations concerning the heat equation: (a) is the solution for $\hat{u}(t) = 0$, (b) is its approximation via POD (non adaptive), (c) is the direct LQR solution computed by MATLAB without POD and, finally, the approximate optimal solution obtained coupling POD and HJB. The approximate value function is computed for $\Delta t = 0.1$ $\Delta x = 0.1$ whereas the optimal trajectory as been obtained with $\Delta t = 0.01$. Test 1, and even Test 2, have been solved in about half an hour of CPU time.

Note that in this example the approximate solution is rather accurate because the regularity of the solution is high due to the diffusion term. Since in the limit the solution tends to the average value the choice of the snapshots will not affect too much the solution, i.e. even with a rough choice of the snapshots will give us a good approximation. The difference between Figs. 2c and 2d is due to the fact that the control space is continuous for Fig. 2c and discrete for Fig. 2d.

Test 2: Heat Equation with No-smooth Initial Data In this section we change the initial condition with a function which is only Lipschitz continuous: $y_0(x) = 1 - |x|$. According to Test 1, we consider the same parameters (see Fig. 3). Riccati's equation has been solved by a MATLAB LQR routine. Thus, we have used the solution given by this routine as the correct solution in order to compare the errors in L^1 and L^2 norm between the reduced Riccati's equation and our approach based



(A)

(B)

(C)

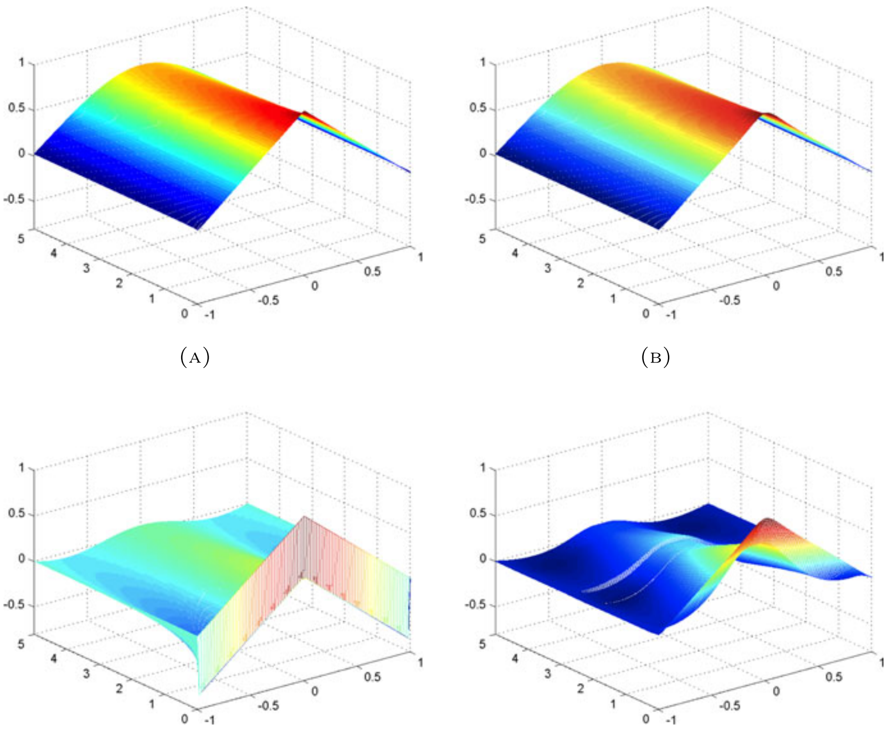
(D)

Fig. 2 Test 1: (**a**) Heat equation without control; (**b**) Heat equation without control, 3 POD basis approximation; (**c**) Controlled solution with LQR-MATLAB; (**d**) Approximate solution POD (3 basis functions) + HJB

on the reduced HJB equation. Since we do not have any information, the snapshots are computed for $\hat{u} = 0$. This is only a guess, but in the parabolic case fits well due to the diffusion term.

As in Test 1, the choice of the snapshots does not effect strongly the approximation due to the asymptotic behavior of the solution. The presence of a Lipschitz continuous initial condition has almost no influence on the global error (see Table 1).

Test 3: Advection-Diffusion Equation The advection-diffusion equation needs a different method. We can not use the same $\hat{y}$ we had in the parabolic case, mainly because in Riccati's equation the control is free and is not bounded, on the contrary when we solve an HJB we have to discretize the space of controls. We modified the problem in order to deal with bang-bang controls. We get $\hat{y}$ in (5.2) just plugging in the control $\hat{u} \equiv 0$. We have considered the control space corresponding only to three values in $[-1, 1]$, then $U = \{-1, 0, 1\}$. We first have tried to get a controlled solution, without any adaptive method and, as expected, we obtained a bad approximation (see Fig. 4).



(A) (B)

(C) (D)

Fig. 3 Test 2: (**a**) Exact solution for $\hat{u} = 0$; (**b**) Exact solution for $\hat{u} = 0$ POD (3 basis functions); (**c**) Approximate optimal solution for LQR-MATLAB; (**d**) Approximate solution POD (3 basis functions) + HJB

Table 1 Test 2: L^1 and L^2 errors at time T for the optimal approximate solution

	L^1	L^2
$y^{LQR} - y^{POD+LQR}$	0.0221	0.0172
$y^{LQR} - y^{POD+HJB}$	0.0204	0.0171

From Fig. 4 it's clear that POD with four basis functions is not able to catch the behavior of the dynamics, so we have applied our adaptive method.

We have consider: $T = 3$, $\Delta x = 0.1$, $\Delta t = 0.008$, $a = -1$, $b = 4$, $R = 0.01$. According to our algorithm, the time interval $[0, 3]$ was divided into $[0, 0.744] \cup [0.744, 1.496] \cup [1.496, 3]$. As we can see our last interval is bigger than the others, this is due to the diffusion term (see Fig. 5). The L^2-error is 0.0761, and the computation of the optimal solution via HJB has required about six hours of CPU time. In Fig. 4 we compare the exact solution with the numerical solution based on a POD representation. Note that, in this case, the choice of only 4 basis functions for the whole interval $[0, T]$ gives a very poor result due to the presence of the advec-

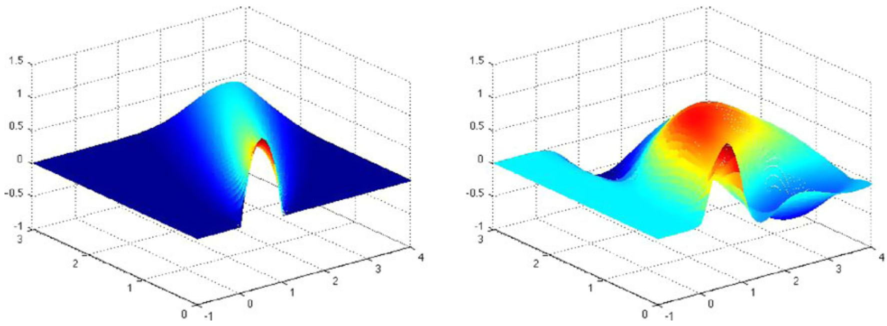


Fig. 4 Test 3: Solution $\hat{y}$ on the *left*, approximate solution on the *right* with POD (4 basis functions)

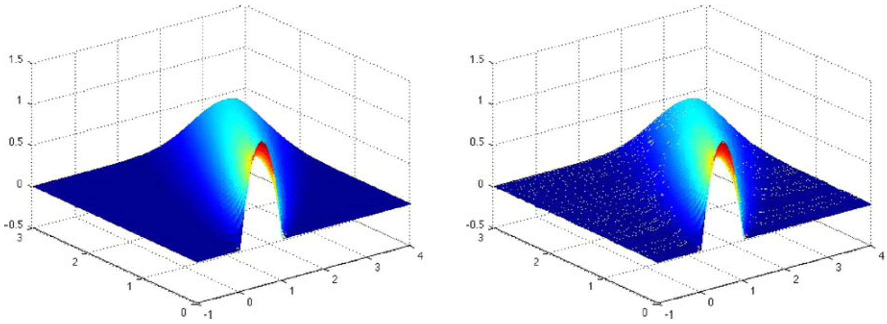


Fig. 5 Test 3: Solution for $\hat{u} \equiv 0$ (*left*), approximate optimal solution (*right*)

tion term. Looking at Fig. 5 one can see the improvement of our adaptive technique which takes always 4 basis functions in each sub-interval.

In order to check the quality of our approximation we have computed the numerical residual, defined as:

$$\mathcal{R}(y) = \left\| y_s(x,s) - \varepsilon y_{xx}(x,s) + c y_x(x,s) - u(s) \right\|.$$

The residual for the solution of the control problem computed without our adaptive technique is 1.1, whereas the residual for the adaptive method is $2 \cdot 10^{-2}$. As expected from the pictures, there is a big difference between these two value.

Test 4: Advection-Diffusion Equation In this test we take a different $\hat{y}$, namely the solution of (5.1) corresponding to the control

$$\hat{u}(t) = \begin{cases} -1 & 0 \le t < 1, \\ 0 & 1 \le t < 2, \\ 1 & 2 \le t \le 3. \end{cases}$$

We want to emphasize we can obtain nice results when the space of controls has few element. The parameters were the same used in Test 3. The L^2-error is 0.09,

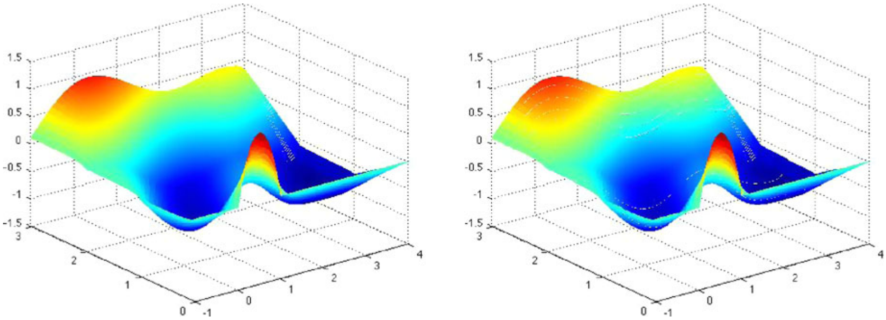


Fig. 6 Test 4: Solution for $\hat{u}$ (*left*), approximate optimal solution (*right*)

and the time was the same we had in Test 3. In Fig. 6 we can see our approximation. In Fig. 6 one can see that the adaptive technique can also deal with discontinuous controls. In this test, the residual for the solution of the control problem without our adaptive technique is 2, whereas the residual for the adaptive method is $3 \cdot 10^{-2}$. Again, the residual shows the higher accuracy of the adaptive routine.

6 Conclusions

As we have discussed, a reasonable coupling between POD and HJB equation can produce feedback controls for infinite dimensional problem. For advection dominated equations that simple idea has to be implemented in a clever way to be successful. It particular, the application of an adaptive technique is crucial to obtain accurate approximations with a low number of POD basis functions. This is still an essential requirement when dealing with the Dynamic Programming approach, which suffers from the curse-of-dimensionality although recent developments in the methods used for HJB equations will allow to increase this bound in the next future (for example by applying patchy techniques).

Another important point is the discretization of the control space. In our examples, the number of optimal control is rather limited and this will be enough for problems which have a bang-bang structure for optimal controls. In general, we will need also an approximation of the control space via reduced basis methods. This point as well as a more detailed analysis of the procedure outlined in this paper will be addressed in our future work.

References

1. M. Bardi, I. Capuzzo Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton–Jacobi-Bellman Equations* (Birkhäuser, Basel, 1997)
2. S. Cacace, E. Cristiani, M. Falcone, A. Picarelli, A patchy dynamic programming scheme for a class of Hamilton–Jacobi-Bellman equations. SIAM J. Sci. Comput. **34**, 2625–2649 (2012)

3. E. Carlini, M. Falcone, R. Ferretti, An efficient algorithm for Hamilton–Jacobi equations in high dimension. Comput. Vis. Sci. **7**(1), 15–29 (2004)
4. M. Falcone, Numerical solution of dynamic programming equations, Appendix in *Optimal Control and Viscosity Solutions of Hamilton–Jacobi-Bellman Equations*, ed. by M. Bardi, I. Capuzzo Dolcetta (Birkhäuser, Boston, 1997), pp. 471–504
5. M. Falcone, R. Ferretti, *Semi-Lagrangian Approximation Schemes for Linear and Hamilton–Jacobi Equations* (SIAM, Philadelphia, to appear)
6. M. Falcone, T. Giorgi, An approximation scheme for evolutive Hamilton–Jacobi equations, in *Stochastic Analysis, Control, Optimization and Applications: A Volume in Honor of W.H. Fleming*, ed. by W.M. McEneaney, G. Yin, Q. Zhang (Birkhäuser, Basel, 1999), pp. 289–303
7. K. Kunisch, S. Volkwein, Control of Burgers' equation by a reduced order approach using proper orthogonal decomposition. J. Optim. Theory Appl. **102**, 345–371 (1999)
8. K. Kunisch, S. Volkwein, Galerkin proper orthogonal decomposition methods for parabolic problems. Numer. Math. **90**, 117–148 (2001)
9. K. Kunisch, S. Volkwein, Optimal snapshot location for computing POD basis functions. ESAIM: M2AN **44**, 509–529 (2010)
10. K. Kunisch, S. Volkwein, L. Xie, HJB-POD based feedback design for the optimal control of evolution problems. SIAM J. Appl. Dyn. Syst. **4**, 701–722 (2004)
11. K. Kunisch, L. Xie, POD-based feedback control of burgers equation by solving the evolutionary HJB equation. Comput. Math. Appl. **49**, 1113–1126 (2005)
12. A.T. Patera, G. Rozza, *Reduced Basis Approximation and a Posteriori Error Estimation for Parameterized Partial Differential Equations*. MIT Pappalardo Graduate Monographs in Mechanical Engineering (2006)
13. M.L. Rapun, J.M. Vega, Reduced order models based on local POD plus Galerkin projection. J. Comput. Phys. **229**, 3046–3063 (2010)
14. F. Tröltzsch, *Optimal Control of Partial Differential Equations: Theory, Methods and Applications* (AMS, Providence, 2010)
15. S. Volkwein, Model reduction using proper orthogonal decomposition (2011). www.math.uni-konstanz.de/numerik/personen/volkwein/index.php

Generalized Sensitivity Analysis for Delay Differential Equations

H.T. Banks, Danielle Robbins, and Karyn L. Sutton

Abstract We present theoretical foundations for traditional sensitivity and generalized sensitivity functions for a general class of nonlinear delay differential equations. Included are theoretical results for sensitivity with respect to the delays. A brief summary of previous results along with several illustrative computational examples are also given.

Keywords Delay equations · Differentiability with respect to delays · Sensitivity · Generalized sensitivity · Fisher information matrix

Mathematics Subject Classification (2010) Primary 34K29 · 34K60 · Secondary 49K40

1 Introduction

Delay differential equations (DDEs) have been used for a number of years to model biological, physical, and sociological processes, as well as other naturally occurring oscillatory systems. Minorsky [56] in 1942 was among the first to introduce the idea of hystero-differential equations, using these type of equations to explain self-excited oscillations arising in dynamic stabilization systems. He proposed [56–58] that some natural phenomena such as self-oscillations may be effected by the

This research was supported in part by Grant Number NIAID R01AI071915-09 from the National Institute of Allergy and Infectious Disease, in part by the US Air Force Office of Scientific Research under grant number FA9550-09-1-0226 and in part by the National Institute on Alcohol Abuse and Alcoholism under a subcontract from the Research Foundation for Mental Hygiene.

H.T. Banks (✉) · D. Robbins · K.L. Sutton
Center for Research in Scientific Computation, Center for Quantitative Science in Biomedicine,
North Carolina State University, Raleigh, NC 27695-8212, USA
e-mail: htbanks@ncsu.edu

D. Robbins
e-mail: danielle.evette.robbins@gmail.com

K.L. Sutton
e-mail: karyn.sutton@gmail.com

K. Bredies et al. (eds.), *Control and Optimization with PDE Constraints*,
International Series of Numerical Mathematics 164,
DOI 10.1007/978-3-0348-0631-2_2, © Springer Basel 2013

previous history of a motion or action as described by a retarded dynamical system. A retarded dynamical system is a system that describes an action that has delayed time dependence. The simplest of these physical systems are usually classified into systems with retarded damping given by

$$\ddot{x}(t) + K\dot{x}(t-\tau) + bx(t) = g(t), \tag{1.1}$$

or those with retarded restoring force described by

$$\ddot{x}(t) + K\dot{x}(t) + bx(t-\tau) = g(t), \tag{1.2}$$

where g is some external force. Specifically, Minorsky used models such as (1.1) and (1.2) to study stabilization systems in ships. It is has been well understood for many years [30] that the infinite degree of the corresponding characteristic equation for a DDE such as (1.1) or (1.2) allows for an infinite number of eigenvalues for even a scalar DDE. This can promote dramatically different (from an ordinary differential equation) solution behavior such as self-excited oscillations in the solution [58]. This property of the DDE along with the widespread presence of delays in many physical and biological systems makes DDEs very important in modeling and control in these systems. Minorsky also provided insight as to the use of a nonlinear DDE to model a system with self-excited oscillations, as a linear DDE is unable to capture all of the properties of the self-excitation. Thus Minorsky lays a foundation for modeling oscillatory phenomena in general systems.

Another early contributor, Hutchinson [47], in 1948 revealed the importance of delay systems in biology and ecology. He developed a delay differential equation model, known as the delay logistic equation, to describe the dynamics of circular causal systems. A circular causal system is any causal system (one with current solution values depending only on current or past inputs) where changes to one part of the system effects another part of the system at a different rate so that the system does not go extinct. Parasite-host interaction is an example of an ecological circular causal system; if a parasite can complete its life cycle without killing the host or drastically altering the growth of the host population, the host population will continue to exist [47, 52]. The delay in this model can represent various naturally occurring phenomena such as the gestation period in a growing population, the life cycle of a parasite, cell cycle delays, etc. Hutchinson's equation (to be used in the numerical illustrations below), its variations and other delay systems have also been used to model physiological control systems as well as numerous other biological processes [1, 3, 15, 16, 24, 26, 33–36, 38–41, 43, 44, 46, 50, 52–55, 61, 62]. This wide spread use of delay equations in applications has continued since the early contributions of Minorsky and Hutchinson.

In the 1970's and 80's much work was done on foundations of delay systems, in contributions both theoretical and qualitative [30, 37, 38, 45] as well as computational (see [4–6, 8, 10–12, 14, 18, 32, 49, 51] and the references therein) in nature. In some of these early efforts, parameter estimation and control system questions led to the investigation of *traditional sensitivity functions* (*TSFs*) for delay systems. These TSFs and a more general concept of *generalized sensitivity functions* (*GSFs*) are the focus of our investigations in this paper. In one early paper [19], Banks, Burns and Cliff observed difficulty when estimating the delay and they suggested that this could be due to the fact that solutions of DDEs may not always be differentiable with respect to the delays; this makes estimation methods such as least squares and maximum likelihood challenging if derivative-based optimization routines are used. These authors also suggested the need for a formal theory regarding the existence of sensitivity functions with respect to the delay. Gibson and Clark [42] and Brewer [31] were among the first to treat theoretical questions of sensitivity for linear DDEs. In both contributions, these authors reformulated the delay system (as was done in the early semigroup approximation efforts of Banks, Burns and Kappel [10–12]) as an abstract system

$$\begin{aligned} \dot{z}(t) &= \mathcal{A}(q)z(t) + \mathcal{B}u(t), \quad t \geq 0, \\ z(0) &= z_0 = (\eta, \phi), \end{aligned} \tag{1.3}$$

where $(\eta, \phi) \in Z \equiv \mathbb{R}^n \times L_2(-r, 0; \mathbb{R}^n)$, $q \in Q$ and the infinitesimal generator $\mathcal{A}(q)$ is defined such that

$$\mathcal{A}(q)\big(\phi(0), \phi\big) = \big(L(q)\phi, \dot{\phi}\big).$$

Then given $t \geq 0$, $S(t; q) : Z \to Z$ is defined such that

$$S(t; q)(\eta, \phi) = \big(x(t; q), x_t(q)\big)$$

where $S(t; q)$ is a strongly continuous semigroup [59] and $x_t(\xi) = x(t + \xi)$, $-r \leq \xi \leq 0$. By defining strongly continuous solution semigroups, a well-posed problem can be formulated. While both contributions present conditions under which solutions are Frechèt differentiable with respect to the parameter q, in Gibson and Clark's efforts [42] the differentiability results were obtained where the operator $\mathcal{A}(q)$ is required to be represented as a linear combination of an operator A independent of the parameter and a dependent *bounded* linear operator $A_1(q)$, i.e., $\mathcal{A}(q) = A + A_1(q)$ with A_1 bounded. Brewer [31] expands the results in [42] by considering classes of problems in which the full parameter dependent operator, $\mathcal{A}(q)$, is unbounded. In Brewer's theory this operator generates a strongly continuous semigroup, and using semigroup representation results he is able to prove the existence of Frechèt derivatives with respect to the parameters for the initial value problem (1.3). As a result of the existence of the Frechèt derivatives, he is able to carefully and rigorously define sensitivity equations with respect to the parameters *including the delay* for the abstract system.

In a more recent report [2], Baker and Rihan *formally* derive sensitivity equations for delay differential equation models, as well as the equations for the sensitivity of parameter estimates with respect to observations (these latter sensitivities are what we shall discuss below as *Generalized Sensitivity Functions* (*GSFs*)). They consider a general nonlinear system of parameter dependent delay differential equations with parameter $p \in \mathbb{R}^L$ given by

$$\begin{aligned} \dot{x}(t,p) &= f\big(t, x(t), x(t-\tau), p\big), \quad t \geq 0, \\ x(t,p) &= \psi(t,p), \quad t \leq 0, \end{aligned} \tag{1.4}$$

and investigate methods for sensitivity functions with respect to the parameters p and delay τ.

Baker and Rihan offer an outline on how to numerically compute both TSFs and GSFs for retarded delay differential equations (as well as for neutral delay differential equations which are not discussed here in any generality). While their focus is on computational methods, they also list issues that arise when carrying out parameter estimation in DDEs. As we have already noted earlier, these include difficulty in establishing existence of the derivatives of the solution with respect to the parameters and the delays, as well as difficulty in establishing well-posedness for the derived sensitivity equations. Some of these issues are dealt with in a rigorous manner below.

Banks and Bortz [9] were among the first to consider sensitivity with respect to distributional delays. They used sensitivity analysis to show how changes in distributed parameters will effect the solutions of their nonlinear delay differential equation model for HIV progression at the cellular level where intracellular processing delays are distributed across cell populations. The models are validated with what is called *aggregate data* [8].

When deriving the sensitivity equations Banks and Bortz obtain a system of DDEs, which are assumed to be well-posed. In their discussion of well-posedness for these sensitivity equations they assume the delay distributions are differentiable and parameterizable by a mean and standard deviation. In [9] they use theoretical steps (i.e., successive approximations, fixed point theory, Lipschitz continuity, etc.) employed in [7] to prove existence and uniqueness of the resulting sensitivities and sensitivity equations. Motivated by the efforts in [9], Banks and Nguyen [17] develop a rigorous theoretical framework for sensitivity functions for general nonlinear dynamical systems in a Banach space X where the parameters μ are themselves members of another Banach space $\mathcal{M}$. In this setting they consider the sensitivity of solutions x with respect to parameters μ in the following type of abstract nonlinear ordinary differential equations

$$\begin{aligned} \dot{x}(t) &= f\big(t, x(t), \mu\big), \quad t \geq t_0, \\ x(t_0) &= x_0, \end{aligned} \tag{1.5}$$

where $f : \mathbb{R}_+ \times X \times \mathcal{M} \to X$ and $\mathcal{M}$ and X are complex Banach spaces. They establish well-posedness for (1.5), and existence of Frechèt derivatives of the solution $x(t)$ with respect to the parameters μ. As a result, there is a unique solution to the corresponding sensitivity equation

$$\begin{aligned} \dot{y}(t) &= f_x\big(t, x(t,t_0,x_0,\mu), \mu\big) y(t) + f_\mu\big(t, x(t,t_0,x_0,\mu), \mu\big), \quad t \geq t_0 \\ y(t_0) &= 0, \end{aligned} \tag{1.6}$$

where $y(t) = \frac{\partial x(t)}{\partial \mu}$. In [17] Banks and Nguyen provide rigorous theoretical sensitivity results for the DDE example for HIV dynamics with measure dependent or distributional parameters given in [9]; however they only present results for the sensitivity with respect to absolutely continuous probability distributions for the delay. In subsequent efforts [22] a rigorous theoretical foundation is developed for sensitivity theory using *directional derivatives* where the parameter space $\mathcal{M}$ is taken as the convex metric space of probability measures (including discrete, continuous or convex combinations thereof) taken with the Prohorov metric topology [8]. Below we give new results for sensitivity with respect to discrete delays. The proofs, given in [28], while quite tedious, continue with an adaption of the well known ideas for existence and uniqueness of the Frechèt derivative with respect to the delay in nonlinear DDE as employed in [9, 17, 22]. Very recent efforts, especially in areas of biology, demonstrate the continuing interest and importance of sensitivity equations in the sciences. For example Burns, Cliff, and Doughty [33] explain the use of continuous sensitivity equations for DDE models arising in a model for Chlamydia Trachomatis, while Kappel [50] discusses generalized sensitivities in dynamics of threshold-driven infections.

After summarizing recent theoretical results on differentiability with respect to parameters, initial conditions and discrete delays, we discuss both traditional and generalized sensitivity functions with respect to the same quantities. Finally, to illustrate computationally the use of these sensitivity functions, we turn to two classical examples: the Hutchinson delayed growth model and the harmonic oscillator with delays.

2 Solutions and Their Approximation

We first summarize some fundamental well-posedness and approximation results that have been recently developed elsewhere [8, 60]. We consider nonlinear nonautonomous dynamical systems involving delays of the general form

$$\begin{aligned} \dot{x}(t) &= G\big(t, x(t), x_t, x(t-\tau_1), \ldots, x(t-\tau_m), \theta\big) + G_2(t), \quad 0 \le t \le T, \\ x(\xi) &= \phi(\xi), \quad -r \le \xi \le 0, \end{aligned} \tag{2.1}$$

where $G = G(t, \eta, \psi, y_1, \ldots, y_m, \theta) : [0, T] \times X \times \mathbb{R}^{nm} \times \mathbb{R}^p \to \mathbb{R}^n$. Here $X = \mathbb{R}^n \times L_2(-r, 0; \mathbb{R}^n)$, $0 < \tau_1 < \cdots < \tau_m = r$, x_t denotes the usual function $x_t(\xi) = x(t+\xi)$, $-r \le \xi \le 0$, and $\phi \in H^1(-r, 0)$. Here $G_2(t)$ is an n-vector input, such as, for example, a control input or in the case of inverse problems a stimulating input to excite the dynamics.

The theoretical results in this manuscript will be illustrated computationally, and therefore, the solutions will need to be approximated. In order to approximate solutions, one may first convert the dynamical system to an abstract evolution equation and then approximate in a space spanned by piecewise linear (or even higher order) splines (i.e., in a Galerkin approach, which is equivalent to a linear finite element approximation in partial differential equations). One is then able to numerically calculate the generalized Fourier coefficients of approximate solutions relative to the splines, and with these coefficients, recover an approximation to the solutions of (2.1).

We turn to the mathematical aspects of these nonlinear functional differential equations (FDE) systems and present an outline of the necessary mathematical and numerical foundations. First we describe the conversion of the nonlinear FDE system to an abstract evolution equation (AEE) as well as provide existence and uniqueness results for a solution to the FDE. One can use the ideas of the linear semigroup framework, in which approximation of linear delay systems has been developed, as a basis for a wide class of nonlinear delay system approximations. Details in this direction can be found in the early work [5, 6, 49] which is a direct extension of the results in [10–12] to nonlinear delay systems. We then provide a fundamental approximation framework including convergence results.

We shall make use of the following hypotheses throughout our presentation.

(H1) The function G satisfies a global Lipschitz condition:

$$\begin{aligned}&\big|G(t,\eta,\psi,y_1,\dots,y_m,\theta)-G(t,\tilde{\eta},\tilde{\psi},w_1,\dots,w_m,\theta)\big|\\&\le K\left(|\eta-\tilde{\eta}|+|\psi-\tilde{\psi}|+\sum_{i=1}^{m}|y_i-w_i|\right)\end{aligned}$$

for some fixed constant K and all $(\eta,\psi,y_1,\dots,y_m)$, $(\tilde{\eta},\tilde{\psi},w_1,\dots,w_m)$ in $X\times\mathbb{R}^{nm}$ uniformly in t and θ.

(H2) The function $G:[0,T]\times X\times\mathbb{R}^{nm}\times\mathbb{R}^p\to\mathbb{R}^n$ is differentiable.

Remark 1 If we define the function $g:[0,T]\times\mathbb{R}^n\times C(-r,0;\mathbb{R}^n)\times\mathbb{R}^p\subset[0,T]\times X\times\mathbb{R}^p\to\mathbb{R}^n$ given by

$$g(t,x,\theta)=g(t,\eta,\psi,\theta)=G\big(t,\eta,\psi,\psi(-\tau_1),\dots,\psi(-\tau_m),\theta\big), \tag{2.2}$$

we observe that even though G satisfies (H1), g will not satisfy a continuity hypothesis on its domain in the X norm.

Letting $z(t)=(x(t),x_t)\in X$, where the Hilbert space X has the inner product

$$\big\langle(\eta,\phi),(\zeta,\psi)\big\rangle_X=\langle\eta,\zeta\rangle_{\mathbb{R}^n}+\int_{-r}^{0}\big\langle\phi(\xi),\psi(\xi)\big\rangle_{\mathbb{R}^n}d\xi, \tag{2.3}$$

we define for each (t,θ) the nonlinear operator $\mathcal{A}(t,\theta):\mathcal{D}(\mathcal{A}(t,\theta))\subset X\to X$ by

$$\begin{aligned}\mathcal{D}\big(\mathcal{A}(t,\theta)\big)&\equiv\big\{\big(\psi(0),\psi\big)\mid\psi\in H^1(-r,0)\big\},\\\mathcal{A}(t,\theta)\big(\psi(0),\psi\big)&=\big(g\big(t,\psi(0),\psi,\theta\big),D\psi\big)\end{aligned}$$

where here $D\psi = \psi'$. Note that $\mathcal{D}(\mathcal{A}(t,\theta))$ does not actually depend on (t,θ) even though the operator does. Then the FDE (2.1) can be formulated as

$$\begin{aligned} \dot{z}(t) &= \mathcal{A}(t)z(t) + G_2(t), \\ z(0) &= z_0, \end{aligned} \tag{2.4}$$

where $z_0 = (x_0, \phi)$ is the initial condition and $\mathcal{A}(t) = \mathcal{A}(t,\theta)$. For notational convenience we suppress the dependence on θ in the remainder of this section.

Theorem 2 *Assume that* (H1) *holds and let* $z(t;\phi,G_2) = (x(t;\phi,G_2), x_t(\phi,G_2))$, *where* x *is the solution of* (2.1) *corresponding to* $\phi \in H^1$, $G_2 \in L_2$. *Then for* $\zeta = (\phi(0),\phi)$, $z(t;\phi,G_2)$ *is the unique solution on* $[0,T]$ *of*

$$z(t) = \zeta + \int_0^t \big[\mathcal{A}(\sigma)z(\sigma) + \big(G_2(\sigma),0\big)\big]d\sigma. \tag{2.5}$$

Furthermore, $G_2 \to z(t;\phi,G_2)$ *is weakly sequentially continuous from* L_2 (*with weak topology*) *to* X (*with strong topology*).

These results can be established in one of several ways including fixed point theorem arguments or Picard iteration arguments. Either of these approaches can be used to establish existence, uniqueness and continuous dependence of the solution of (2.5). For existence, uniqueness and continuous dependence of the solution of (2.1), we note that our condition (H1) is a global version of the hypothesis of Kappel and Schappacher in [51], so that in the autonomous case their results also yield immediately the desired result for (2.1).

The uniqueness of solutions to (2.5) follows in the usual manner once we establish that $\mathcal{A}$ satisfies a dissipative inequality. Indeed, we define a weighting function w on $[-r,0)$ by

$$w(\xi) = j \quad \text{for } \xi \in [-\tau_{m-j+1}, -\tau_{m-j}),\ j = 1,2,\ldots,m.$$

Then we consider solutions on the space X_w which is topologically equivalent to X, with the weighted inner product

$$\big\langle(\eta,\phi),(\zeta,\psi)\big\rangle_w = \langle\eta,\zeta\rangle_{\mathbb{R}^n} + \int_{-r}^0 \big\langle\phi(\xi),\psi(\xi)\big\rangle_{\mathbb{R}^n} w(\xi)d\xi. \tag{2.6}$$

One can show without difficulty that (H1) implies the dissipative inequality (see [29, p. 71]) for the nonlinear operator $\mathcal{A}(t)$

$$\big\langle\mathcal{A}(t)x - \mathcal{A}(t)y, x-y\big\rangle_w \le \omega\langle x-y, x-y\rangle_w \tag{2.7}$$

for all $x, y \in \mathcal{D}(\mathcal{A}(t))$ and all t.

The system of functional differential or delay equations described can now be simulated using an algorithm first developed by Banks and Kappel for linear systems [12] and extended in [5, 6]. Solutions are approximated in a space spanned by piecewise linear splines. Thus one can numerically calculate the generalized Fourier coefficients of the approximate solution in the spline basis representation and recover an approximation to the solution of (2.1).

Define X^N to be an approximating subspace [12, 13] of X. In particular, we choose $X^N = X_1^N$ to be the piecewise linear spline subspaces of X discussed in detail in [12]. We briefly outline the results for the piecewise linear subspaces X_1^N (see Sect. 4 of [12]) given by

$$X_1^N = \big\{\big(\phi(0), \phi\big) \mid \phi \text{ is a continuous first-order spline function} \\ \text{with knots at } t_j^N = -jr/N, \, j = 0, 1, \ldots, N\big\}.$$

A careful study of the arguments behind our presentation reveals that the approximation results given here hold for more general spline approximations. For example, if one were to treat cubic spline approximations (X_3^N of [12]), one would use the appropriate approximation analogues of Theorem 2.5 of [63] and Theorem 21 of [64] (e.g., see Theorem 4.5 of [63]). Hereafter when we write X^N the reader should understand that we mean X_1^N of [12].

Let P^N be the orthogonal projection in $\langle\cdot,\cdot\rangle_w$ of $X = X_w$ onto X^N. We define the approximating operator $\mathcal{A}^N(t) = P^N\mathcal{A}(t)P^N$ and consider the approximating equations in X^N given by

$$z^N(t) = P^N\zeta + \int_0^t \big[\mathcal{A}^N(\sigma)x^N(\sigma) + P^N\big(G_2(\sigma), 0\big)\big]d\sigma. \tag{2.8}$$

These are equivalent to

$$\dot{z}^N(t) = \mathcal{A}^N(t)z^N(t) + P^N\big(G_2(t), 0\big), \qquad z^N(0) = P^N\zeta, \tag{2.9}$$

the finite dimensional system in X^N.

Define $\alpha^N(t)$ so that $x^N(t) = \hat{\beta}^N\alpha^N(t)$ for any $x^N \in X^N$. Here

$$\hat{\beta}^N = \big(\beta^N(0), \beta^N\big) \quad \text{where } \beta^N = \big(e_0^N, e_1^N, \ldots, e_N^N\big).$$

The basis elements e_j^N's are piecewise linear splines defined by the Kronecker symbol δ_{ij}, so

$$e_j^N(t_i) = \delta_{ij} \quad \text{for } i, j = 0, 1, \ldots, N.$$

Then solving for $z^N(t)$ in the finite dimensional system (2.9) is equivalent to solving for $\alpha^N(t)$ in the vector system

$$\begin{aligned} \dot{\alpha}^N(t) &= A^N\alpha^N(t) + P^N G_2(t), \\ \alpha^N(0) &= \alpha_0^N, \end{aligned} \tag{2.10}$$

where $\hat{\beta}^N\alpha_0^N = P^N z_0$ and A^N is the matrix representation for $\mathcal{A}^N$. We note that having obtained $\alpha^N(t)$, the product $\hat{\beta}^N\alpha^N(t)$ converges uniformly in t to the solution $z(t) = (x(t), x_t)$ of (2.4) if we can argue the convergence $z^N(t) \to z(t)$. To do this for linear systems, one can use the Trotter-Kato theorem, involving linear semigroups. For nonlinear autonomous systems, one can invoke the use of nonlinear semigroups [49, 51].

From (2.7) and the definition of $\mathcal{A}^N$ in terms of the self-adjoint projections P^N, we have at once that under (H1) the sequence $\{\mathcal{A}^N\}$ satisfies on X a uniform dissipative inequality

$$\left\langle \mathcal{A}^N(t)x - \mathcal{A}^N(t)y, x - y \right\rangle_w \leq \omega \langle x - y, x - y \rangle_w \tag{2.11}$$

for all $x, y \in \mathcal{D}(\mathcal{A}(t))$ and all t. We note that ω does not depend on N. Uniqueness of solutions of (2.8) then follows immediately from this inequality. Upon recognition that (2.9) is equivalent to a nonlinear ordinary differential equation in Euclidean space with the right-hand side satisfying a global Lipschitz condition, one can easily argue existence of solutions for (2.9) and hence for (2.8) on any finite interval $[0, T]$. The next theorem, which ensures that solutions of (2.9) converge to those of (2.5), along with its proof is contained in [26].

Theorem 3 *Assume* (*H*1), (*H*2). *Let* $\zeta = (\phi(0), \phi)$, $\phi \in H^1$ *and* $G_2 \in L_2(0, T)$ *be given*, *with* z^N *and* z *the corresponding solutions on* $[0, T]$ *of* (2.9) *and* (2.5), *respectively. Then* $z^N(t) \to z(t) = (x(t; \phi, G_2), x_t(\phi, G_2))$, *as* $N \to \infty$, *uniformly in* t *on* $[0, T]$.

Remark 4 One can actually obtain slightly stronger results than those given in Theorem 3. One can consider solutions of (2.5) and (2.9) corresponding to initial data $z_0 = (x_0, \phi) = \zeta$ with $x_0 \in \mathbb{R}^n$, $\phi \in L_2$ (i.e., $\zeta \in X$) and argue that the results of Theorem 3 hold also in this case.

The convergence given in Theorem 3 yields state approximation techniques for nonlinear FDE systems based on the spline methods developed in [12]. These results can be applied directly to control and identification problems, which are discussed in [5, 6].

3 Continuous Dependence and Differentiability

To establish continuous dependence in parameters and differentiability with respect to model parameters, initial conditions, and discrete time delays (not previously done elsewhere for general nonlinear systems to the authors' knowledge), we focus on a restricted case with nonlinear autonomous systems with one discrete delay of the form

$$\frac{dx(t)}{dt} = G\big(x(t), x(t - \tau), \theta\big), \quad t > 0, \tag{3.1}$$

$$x(\xi) = \begin{cases} \phi(\xi), & -\tau \leq \xi < 0 \\ x_0, & \xi = 0 \end{cases} \tag{3.2}$$

where $x(t) \in \mathbb{R}^n$, $x_0 \in \mathbb{R}^n$, and $\theta \in \mathbb{R}^p$. While we consider here the case of finite dimensional model parameters the results also hold in a more general case when parameters are distributed, and hence infinite dimensional, as presented in [21, 22]. Once established, these results allow us to study the traditional and generalized sensitivity functions, where sensitivity is considered with respect to these three quantities. We begin by considering continuous dependence of solutions $x(t)$ on model parameters θ. We note that we focus on these quantities as they are often unknown in practice and may need to be estimated from observed or experimental data. The use of sensitivity functions can aid in that endeavor.

Lemma 5 *Let $G : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n$ and for $\theta = \theta_0$, let $x(t, x_0, \phi, \tau, \theta_0)$ be a solution of (3.1)–(3.2) for $t \in [0, T]$. Assume that*

$$\lim_{\theta \to \theta_0} G(x, \tilde{x}, \theta) = G(x, \tilde{x}, \theta_0), \tag{3.3}$$

uniformly in x and $\tilde{x}$. For $(x_1, \tilde{x}_1, \theta), (x_2, \tilde{x}_2, \theta) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^p$ assume that

$$\left| G(x_1, \tilde{x_1}, \theta) - G(x_2, \tilde{x_2}, \theta) \right| \le C_1 |x_1 - x_2| + C_2 |\tilde{x_1} - \tilde{x_2}| \tag{3.4}$$

where $C_j > 0$ is a constant for $j = 1, 2$. Then the initial value problem (3.1)–(3.2) has a unique solution $x(t, x_0, \phi, \tau, \theta)$ that satisfies

$$\lim_{\theta \to \theta_0} x(t, x_0, \phi, \tau, \theta) = x(t, x_0, \phi, \tau, \theta_0), \quad t \in [0, T].$$

Next, we turn to differentiability of the general system (3.1)–(3.2) with respect to model parameters in the following theorem. The proof is excluded but can be found in [28, 60]. Also, the proof of Theorem 8 can easily be followed to give that of Theorem 6. Without further discussion, we then state Theorem 7, in which we establish differentiability of the model system with respect to the initial conditions, which is also proven in [28].

Theorem 6 *Suppose that $G(x, \tilde{x}, \theta)$ has continuous Frechèt derivatives G_θ, G_x, $G_{\tilde{x}}$ such that $|G_x| \le M_0$, $|G_{\tilde{x}}| \le M_1$, and $|G_\theta| \le M_2$. Then the Frechèt derivative $y_1(t) = \frac{\partial x(t)}{\partial \theta} \in \mathbb{R}^{n \times p}$ exists and is the unique solution for*

$$\begin{aligned} \dot{y}_1(t) &= G_x\big(x(t), x(t-\tau), \theta\big) y_1(t) + G_{\tilde{x}}\big(x(t), x(t-\tau), \theta\big) y_1(t-\tau) \\ &\quad + G_\theta\big(x(t), x(t-\tau), \theta\big), \\ y_1(t) &= 0, \quad -\tau \le t < 0 \end{aligned} \tag{3.5}$$

Theorem 7 *Suppose the function $G(x, \tilde{x}, \theta)$ of (3.1) has continuous Frechèt derivatives $G_x(x, \tilde{x}, \theta)$, $G_{\tilde{x}}(x, \tilde{x}, \theta)$, with respect to x and $\tilde{x}$, with $|G_x| \le M_0$, $|G_{\tilde{x}}| \le M_1$. Then the Frechèt derivative $y_2(t) = \frac{\partial}{\partial z_0} x(t, z_0, \theta)$ exists with $y_2(t) \in \mathcal{L}(Z, \mathbb{R}^n)$ (recall $z_0 = (x_0, \phi)$, $Z = \mathbb{R}^n \times L^2(-\tau, 0; \mathbb{R}^n)$), and satisfies the equation*

$$\begin{aligned}\dot{y}_2(t)[h] &= G_x\big(x(t), x(t-\tau), \theta\big) y_2(t)[h] \\ &\quad + G_{\tilde{x}}\big(x(t), x(t-\tau), \theta\big) y_2(t-\tau)[h], \quad t > 0, \qquad (3.6)\\ y_2(\xi) &= \mathcal{I}, \quad -\tau \le \xi \le 0,\end{aligned}$$

where $\mathcal{I} \in \mathcal{L}(Z, \mathbb{R}^n)$ is the identity.

Finally we state results for derivatives with respect to the discrete delays with proofs being given in [28].

Theorem 8 *Suppose that $G(x, \tilde{x}, \theta)$ has continuous Frechèt derivatives G_x, $G_{\tilde{x}}$ such that $|G_x| \le M_0$, and $|G_{\tilde{x}}| \le M_1$ and suppose that the solution x of (3.1)–(3.2) satisfies $x \in H^{1,\infty}(-\tau, T; \mathbb{R}^n)$, for $0 < \tau < r$ for fixed $r > 0$. Then the Frechèt derivative $y_3(t) = \frac{\partial x(t)}{\partial \tau} \in \mathbb{R}^n$ exists and is the unique solution for*

$$\begin{aligned}\dot{y}_3(t) &= G_x\big(x(t), x(t-\tau), \theta\big) y_3(t) \\ &\quad + G_{\tilde{x}}\big(x(t), x(t-\tau), \theta\big)\big[y_3(t-\tau) - \dot{x}(t-\tau)\big], \qquad (3.7)\\ y_3(\xi) &= 0, \quad -\tau \le \xi \le 0. \qquad (3.8)\end{aligned}$$

Moreover, $\frac{\partial x(t)}{\partial \tau}$ is continuous in θ and, if $x \in C^1(-\tau, T; \mathbb{R}^n)$ it is also continuous in τ.

4 Sensitivity Functions

Given the above results, especially differentiability in the quantities θ, $z_0 = (x_0, \phi)$ and τ, we are now able to use the powerful sensitivity analytic techniques in delay systems. For further simplification in the remainder of our discussions we restrict our considerations to constant function initial conditions so in $z_0 = (x_0, \phi)$ we assume $\phi(\xi) = x_0$, $-r \le \xi \le 0$. Traditionally, sensitivity analysis is the quantification of the effect changes in parameters have on model solutions. Traditional sensitivity functions (TSFs), which are given by,

$$\begin{aligned}y_1^k(t) &= \frac{\partial x}{\partial \theta^k} \quad k = 1, \ldots, p,\\ y_2^l(t) &= \frac{\partial x}{\partial x_0^l} \quad l = 1, \ldots, n, \qquad (4.1)\\ y_3(t) &= \frac{\partial x}{\partial \tau},\end{aligned}$$

are local in nature as they are defined by locally evaluated partial derivatives, i.e., $\frac{\partial x}{\partial \theta}(t, \bar{\theta}, \bar{x}_0, \bar{\tau})$, which gives information over specified time intervals, and at values of parameters, initial conditions and delays. Even with this limitation, these functions have been used to improve sampling in an experimental setting; specifically they can be used to guide the time at which measurements should be taken to best inform the estimation of unknown parameters [20, 25]. That is, sampling might be advisable in time intervals where, for example, $y_1^k(t)$ is large, as it indicates that the model solution $x(t)$ is sensitive to changes in the parameter θ_k. Similarly, insensitivity to a certain parameter (or unknown quantity), indicated by small or zero values of the TSF, imply that observations can not be profitably taken in that region if the goal is estimation of the parameter.

TSFs may be approximated by forward differences, but are typically found by solving the system of sensitivity equations

$$\frac{d}{dt}\frac{\partial x(t)}{\partial \theta} = \frac{\partial G}{\partial x}\frac{\partial x}{\partial \theta}(t) + \frac{\partial G}{\partial \tilde{x}}\frac{\partial x}{\partial \theta}(t-\tau) + \frac{\partial G}{\partial \theta}(t) \tag{4.2}$$

for the corresponding system

$$\begin{aligned} \frac{dx(t)}{dt} &= G\big(x(t), x(t-\tau), \theta\big), \quad t > 0, \\ x(\xi) &= x_0, \quad -r \le \xi \le 0, \end{aligned} \tag{4.3}$$

where the $\frac{\partial}{\partial \theta}$ and $\frac{d}{dt}$ operators have been interchanged, due to the continuity assumptions made on G and x. We note that sensitivity analysis is most efficiently carried out in two steps. Once a solution $x(t)$ corresponding to $(\bar{\theta}, \bar{x}_0, \bar{\tau})$ of the above (original delay) equation (4.3) is obtained, one uses this solution to evaluate the coefficients in system (4.2). This decoupling of the original equation and the sensitivity equation has implications when considering the sensitivity with respect to the time delay τ in one of the examples discussed below, which if solved in a coupled manner would result in a so-called neutral delay system.

Generalized sensitivity functions, first introduced by Thomseth and Cobelli [65], and further studied in a series of papers by Banks, et al., [20, 25, 27], provide a measure of how informative measurements of the output or observation variables ($f(t,q)$ defined below, which are not necessarily simply the state variables), are for the identification of unknown quantities. Notably, the functions G and h (the observation operator introduced below) must be differentiable to construct the TSFs, and must also be sufficiently smooth to construct generalized sensitivity functions (GSFs). Before defining the GSFs we briefly outline an inverse problem framework, not only to put our discussion in context, but also to define quantities in the definition of the GSFs.

Given a model solution $x(t)$, the sensitivity of the solution to an estimated quantity q_k (where $q = (\theta, x_0, \tau)^T$) is

$$s_k(t,q) = \frac{\partial f}{\partial q_k}(t,q) \in \mathbb{R}^m,$$

where $f(t,q) = h(t, x(t), x(t-\tau), \theta)$ are the model quantities corresponding to the observed data. Here h is a general observation operator commonly found in inverse problems that detail the type of data being collected (see e.g., [25]). Observations are typically available at discrete times, which we denote by $t_1, \dots, t_{n_d}$. The model representation of the data is then

$$f(t_j, q) = h\bigl(t_j, x(t_j), x(t_j - \tau), \theta\bigr), \quad j = 1, \dots, n_d.$$

In general, the data are not exactly $f(t_j, q)$, due to uncertainty in the measurement process, and also due to small fluctuations not explicitly included in the model. Therefore we represent the observation process Y_j at time t_j by the statistical model

$$Y_j = f\bigl(t_j, q^0\bigr) + \mathcal{E}_j, \quad j = 1, \dots, n_d, \tag{4.4}$$

where $f(t_j, q) = h(t_j, x(t_j), x(t_j - \tau), \theta)$, $q = (\theta, x_0, \tau)$, for $q \in \mathscr{Q} = \mathbb{R}^p \times \mathbb{R}^n \times \mathbb{R}^1$. Here $q^0 = (\theta^0, x_0^0, \tau^0)^T$ represents the 'true values' of the parameters that generates the observations $\{Y_j\}_{j=1}^{n_d}$. The existence of q^0 is commonly assumed [23], implying that (3.1) describes the biological, sociological, or physical process essentially precisely.

The observation errors $\mathcal{E}_j$ are random variables, each with unknown but assumed independent and identical probability distributions of mean zero, and constant variance σ^2. Each data set $\{y_j\}_{j=1}^{n_d}$ is one realization of the random variable $\{Y_j\}_{j=1}^{n_d}$, and the corresponding errors are also a realization of the $\mathcal{E}_j$. Estimating unknown quantities via the minimization between the model and data assuming the statistical model (4.4) gives rise to the commonly used ordinary least squares (OLS) estimator

$$q_{OLS} = \arg\min_{q \in \mathscr{Q}} \sum_{j=1}^{n_d} \bigl| Y_j - f(t_j, q) \bigr|^2, \tag{4.5}$$

where the objective functional is minimized over an admissible parameter space $\mathscr{Q}$. Another common formulation is a weighted least squares procedure, in which the error is assumed to be proportional to the model quantity $f(t_j; q)$, i.e., relative error. For a more complete discussion of the underlying assumptions and related formulations, see [23].

The variance σ^2 of the observation error is used in the computation of standard errors, confidence intervals, etc. and also in the generalized sensitivity functions. For a given set of data, $\{y_j\}_{j=1}^{n_d}$ and parameter estimates $\hat{q}$, the (bias-adjusted) variance is estimated as

$$\hat{\sigma}^2 = \frac{1}{n_d - n_p} \sum_{j=1}^{n_d} \bigl| y_j - f(t_j, \hat{q}) \bigr|^2 \tag{4.6}$$

for $n_p = p + n + 1$ estimated parameters, where $n_p = \dim(\mathscr{Q})$.

The generalized sensitivity functions [25, 27, 65] are defined by

$$\mathbf{gs}(t) = \int_0^t \left[F(T)^{-1} \frac{1}{\sigma^2(s)} \nabla_q f\bigl(s, q^0\bigr) \right] \cdot \nabla_q f\bigl(s, q^0\bigr) dP(s), \quad t \in [0, T], \tag{4.7}$$

for variance $\sigma^2(t)$ that may possibly be time-dependent, true parameters q^0, some general measure P that embodies the observations, and the Fisher information matrix (FIM) F which is defined by

$$F(T) = \int_0^T \frac{1}{\sigma^2(t)} \nabla_q f\big(t, q^0\big) \nabla_q f\big(t, q^0\big)^T dP(t). \tag{4.8}$$

We note that the definition of the measure P affects the FIM, and it can be chosen in such a way as to optimize the information from data concerning the estimated parameters. The GSFs are cumulative functions, such that at time t_j, only the contributions of measurements up to and including those at time t_j are relevant. By the definition in (4.7), it is readily seen that the GSFs are one at the final time $\mathbf{gs}(T) = 1$. As discussed in [25, 65], regions which contain the *sharpest change* (either increase or decrease) of the GSFs are regions of high information content. Decreases in the GSF corresponding to a given parameter indicate correlation between that parameter and at least one other estimated parameter. In this case, it can be seen [25] that computing the GSF for one of the correlated parameters and holding the other(s) fixed, will result in a monotonically increasing GSF. Therefore, regions over which the GSF decreases indicate that the data in that region indeed contains information concerning that parameter, but it is correlated with at least one other parameter, and simultaneous identifiability of all parameters may be difficult.

As observations are typically available at discrete time points and our discussions are in the context of parameter estimation from observed or measured data, we have included here also the definitions for the GSFs and FIM for a discrete measure $P = \sum_{j=1}^{n_d} \Delta_{t_j}$. In the discrete case, the generalized sensitivity functions are

$$\mathbf{gs}(t_j) = \sum_{i=1}^{j} \frac{1}{\sigma^2(t_j)} \big[F^{-1} \times \nabla_q f\big(t_i, q^0\big)\big] \cdot \nabla_q f\big(t_i, q^0\big), \tag{4.9}$$

for observation times t_j where $j = 1, \dots, n_d$. In the above definition, the discrete FIM is given by

$$F = \sum_{j=1}^{n_d} \frac{1}{\sigma^2(t_j)} \nabla_q f\big(t_j, q^0\big) \nabla_q f\big(t_j, q^0\big)^T, \tag{4.10}$$

which measures the information content of the data corresponding to the parameters. In both (4.7) and (4.9), the (biased) estimate for the variance of the observation error is used up to and including the time t_j of the observation, given by

$$\sigma^2(t_j) = \frac{1}{j} \sum_{i=1}^{j} \big|y_i - f(t_i, \hat{q})\big|^2. \tag{4.11}$$

If the variance is assumed constant ($\sigma^2(t) \equiv \sigma^2$), one would simply calculate the estimate as in (4.6), and use that in (4.7) or (4.9).

5 Illustrative Computations

To complete our presentation, we illustrate the uses of sensitivity analysis in two prominent examples of delay equations. The first example we consider is a delay version of one of the most commonly studied models, the logistic equation. This delayed logistic equation, commonly known as Hutchinson's equation mentioned above, is not only discussed in most introductory modeling courses, but is still used in research endeavors to represent growth within an environment in which saturation is possible, but the death rate is proportional to previous population levels. The standard logistic example (without delay) has been effectively used to illustrate with simulated data the ideas of traditional and sensitivity functions and how these techniques may improve data sampling for the purpose of parameter estimation [20, 25]. Therefore, it is natural to turn to the delayed logistic equation now that we are able to study sensitivity functions in systems involving a discrete delay. Here we will also numerically generate simulated data with a known delay, and demonstrate that the estimation can be improved using insights gained from the sensitivity function solutions.

The second example we use is also an ubiquitous model, the delayed harmonic oscillator of Minorsky discussed in the Introduction. As noted there, this example arises in many physical applications where oscillatory phenomena are important.

5.1 Hutchinson Equation Example

In his seminal paper [47] and book [48], Hutchinson arrived at a version of the logistic equation that incorporated a delay in the carrying or death rate term,

$$\frac{dx(t)}{dt} = rx(t)\left(1 - \frac{x(t-\tau)}{K}\right). \tag{5.1}$$

The model was suggested as a possible explanation of the growth dynamics seen in Daphnia. This population seemed to grow exponentially at low population sizes, but it would oscillate at higher population levels. Hutchinson hypothesized that this growth was like that of the logistic model, only that the population seemed to be able to exceed its carrying capacity and perhaps it was this value that the population level was oscillating around.

The traditional sensitivity functions with respect to the model parameters r, K, initial condition x_0, and delay τ are given by

$$\frac{\partial}{\partial r}\frac{dx(t)}{dt} = r\left[1 - \frac{x(t-\tau)}{K}\right]\frac{\partial x(t)}{\partial r} - \frac{rx(t)}{K}\frac{\partial x(t-\tau)}{\partial r} + x(t)\left[1 - \frac{x(t-\tau)}{K}\right],$$

$$\frac{\partial}{\partial K}\frac{dx(t)}{dt} = r\left[1 - \frac{x(t-\tau)}{K}\right]\frac{\partial x(t)}{\partial K} - \frac{rx(t)}{K}\frac{\partial x(t-\tau)}{\partial K} + rx(t)\left[\frac{x(t-\tau)}{K^2}\right],$$

$$\frac{\partial}{\partial x_0}\frac{dx(t)}{dt} = r\left[1-\frac{x(t-\tau)}{K}\right]\frac{\partial x(t)}{\partial x_0} - \frac{rx(t)}{K}\frac{\partial x(t-\tau)}{\partial x_0},$$

$$\frac{\partial}{\partial \tau}\frac{dx(t)}{dt} = r\left[1-\frac{x(t-\tau)}{K}\right]\frac{\partial x(t)}{\partial \tau} - \frac{rx(t)}{K}\left[\frac{\partial x(t-\tau)}{\partial \tau} - \dot{x}(t-\tau)\right].$$

By changing the order of differentiation, and letting $s_1(t) = \frac{\partial x(t)}{\partial r}$, $s_2(t) = \frac{\partial x(t)}{\partial K}$, $s_3(t) = \frac{\partial x(t)}{\partial x_0}$, and $s_4(t) = \frac{\partial x(t)}{\partial \tau}$, we have the system

$$\frac{\partial s_1(t)}{\partial t} = r\left[1-\frac{x(t-\tau)}{K}\right]s_1(t) - \frac{rx(t)}{K}s_1(t-\tau) + x(t)\left[1-\frac{x(t-\tau)}{K}\right], \tag{5.2}$$

$$\frac{\partial s_2(t)}{\partial t} = r\left[1-\frac{x(t-\tau)}{K}\right]s_2(t) - \frac{rx(t)}{K}s_2(t-\tau) + rx(t)\left[\frac{x(t-\tau)}{K^2}\right], \tag{5.3}$$

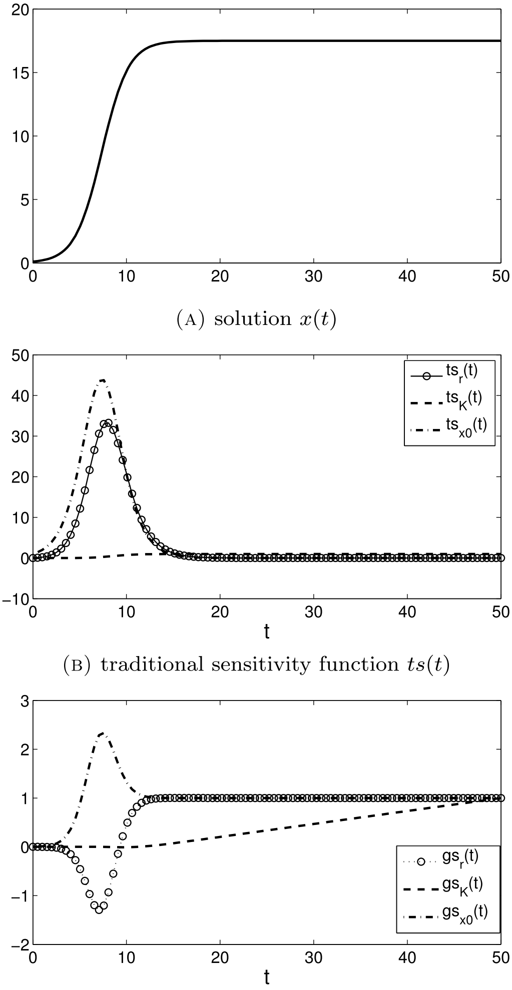
$$\frac{\partial s_3(t)}{\partial t} = r\left[1-\frac{x(t-\tau)}{K}\right]s_3(t) - \frac{rx(t)}{K}s_3(t-\tau), \tag{5.4}$$

$$\frac{\partial s_4(t)}{\partial t} = r\left[1-\frac{x(t-\tau)}{K}\right]s_4(t) - \frac{rx(t)}{K}\left[s_4(t-\tau) - \dot{x}(t-\tau)\right]. \tag{5.5}$$

As noted earlier, we consider only the case of constant initial data, and thus we do not discuss here the Frechèt derivative $y_2(t) = \frac{\partial}{\partial z_0}x(t, z_0, \theta)$ where $z_0 = (x_0, \phi)$, $Z = \mathbb{R}^n \times L^2(\tau, 0; \mathbb{R}^n)$; the results of Theorem 7 still ensure the existence and uniqueness of the solution $\frac{\partial x(t)}{\partial x_0}$ to (5.4), for this simpler case. The existence of unique solutions to (5.2) and (5.3) are guaranteed by Theorem 6, and a unique solution for (5.5) by Theorem 8. Note that (5.5) is not a neutral equation if one assumes the solution $x(t)$ (and also $x(t-\tau)$) is already computed when sensitivity analysis is done; i.e., we decouple the original equation and first solve the delay equation before computing sensitivities. Therefore, when computing sensitivities the $x(t)$ and $x(t-\tau)$ are not unknown quantities but rather an input in the traditional sensitivity functions above.

The solutions for the Hutchinson equation with no delay (i.e., the standard logistic equation), and the corresponding traditional and generalized sensitivity functions are displayed in Fig. 1. In comparing panels Fig. 1(b) to Fig. 1(a), the traditional sensitivity functions with respect to the growth rate r and the initial condition x_0 suggest that the beginning growth portion of the solution is quite sensitive to both parameters. In the bottom panel Fig. 1(c), the solutions of the generalized sensitivity function suggest that the same region is informative for both parameters, but that they are correlated since one of the curves decreases as the other increases. Thus, estimating both the initial condition x_0 and the growth rate r simultaneously from data corresponding to this interval is likely problematic. As one would expect the solution appears to be sensitive to the carrying capacity essentially once it is approached. It is easier to see this in panel Fig. 1(c) than in Fig. 1(b), as the magnitude of the sensitivity to the other parameters (r and x_0) is significantly greater. The definition of the generalized sensitivity functions is such that their magnitude is not as varied even with respect to different quantities.

Fig. 1 The numerical approximation to the solutions (**a**) for the Hutchinson equation, and the corresponding traditional (**b**) and generalized (**c**) sensitivity functions with respect to the model parameters r and K, and the constant initial value x_0 are provided here for the values $\bar{r} = 0.7$, $\bar{K} = 17.5$, and $\bar{x}_0 = 0.1$. The generalized sensitivity functions were computed with constant variance $\sigma^2 = 0.1$

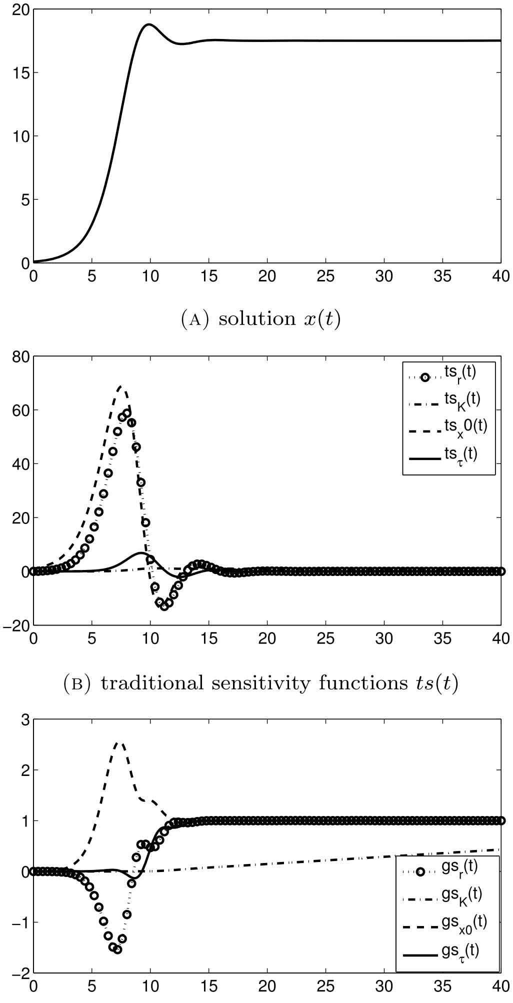


(A) solution $x(t)$

(B) traditional sensitivity function $ts(t)$

(C) generalized sensitivity function $gs(t)$

With a moderate delay, $\tau = 1$, there appears only one time interval over which the solution $x(t)$ exceeds its carrying capacity, as seen in Fig. 2(a). The solution then decreases to below its carrying capacity but the effect is not sufficient for the oscillations to continue, and the solution approaches its carrying capacity $x(t) \to K$ around $t = 14$. It is around the time of the solution first exceeding and then decreasing to less than the carrying capacity (approximately, the interval $t \in [8, 14]$), which can be interpreted as the effect of the delay, that the sensitivity function solutions can be interpreted to mean that the model solution $x(t)$ is sensitive to this delay τ. The solutions of the traditional and generalized sensitivity functions with respect to x_0 and r together suggest that the beginning time interval of the solution is most sensitive to these quantities but that they are strongly correlated.



(a) solution $x(t)$

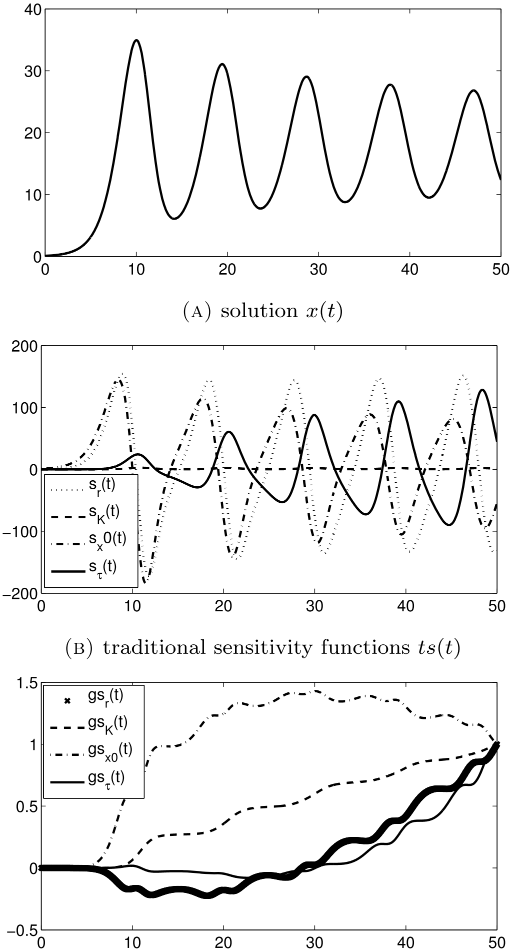
(b) traditional sensitivity functions $ts(t)$

(c) generalized sensitivity functions $gs(t)$

Fig. 2 The numerical approximation for the solutions (**a**) to the Hutchinson equation with delay $\tau = 1$, and corresponding traditional (**b**) and generalized (**c**) sensitivity functions with respect to growth rate r, carrying capacity K, constant initial state x_0, and delay τ, each evaluated at $(\bar{r}, \bar{K}, \bar{x}_0, \bar{\tau}) = (0.7, 17.5, 0.1, 1)$. The generalized sensitivity functions were computed with constant variance $\sigma^2 = 0.1$

With a larger delay, $\tau = \frac{\pi}{2r} \approx 2.244$, the results given in Fig. 3 reveal that many more oscillations in the solution $x(t)$ occur, although they do dampen slightly. The traditional and generalized sensitivity functions for the unknown quantities $q = (r, K, x_0, \tau)$ then indicate which parts of the oscillatory solution are most sensitive to the respective parameter q_i. Regions of decreasing GSF indicate correlation among parameters, as with the growth rate r and initial condition x_0 in Figs. 1 and 2.

To illustrate the information gained from the solutions of the TSF and GSF with respect to the delay with a moderate delay $\tau = 1$, we generated simulated data with 10 % error and used this to estimate the delay τ, while holding the other parameters fixed. As seen in [25], any parameter correlation issues would be irrelevant and



(A) solution $x(t)$

(B) traditional sensitivity functions $ts(t)$

(C) generalized sensitivity functions $gs(t)$

Fig. 3 The numerical approximation for the solutions (**a**) to the Hutchinson equation with delay $\tau = \frac{\pi}{2r} \approx 2.244$, and corresponding traditional (**b**) and generalized (**c**) sensitivity functions with respect to growth rate r, carrying capacity K, constant initial state x_0, and delay τ each evaluated at $(\bar{r}, \bar{K}, \bar{x}_0, \bar{\tau}) = (0.7, 17.5, 0.1, \frac{\pi}{2\bar{r}})$. The generalized sensitivity functions were computed with constant variance $\sigma^2 = 0.1$

estimates should be improved if data is concentrated in any regions of enhanced information content (regions of greatest change in GSF or TSF).

The results from estimating the delay τ from $\{y_j^{unif}\}_{j=1}^{10}$, 10 data points spread uniformly over the time interval [0, 14], versus $\{y_j^{GSF}\}_{j=1}^{10}$ when 8 out of 10 data points are concentrated in the interval [8, 14] are contained in Table 1. Improvement using data with enhanced information content with respect to the delay is evident in that the estimated value for $\hat{\tau}$ is closer to its true value of $\tau = 1$ when the initial guess for $\hat{\tau}$ is either above ($\hat{\tau} = 1.2$) or below ($\hat{\tau} = 0.8$) the true value. Additionally the corresponding standard errors are lower for delays estimated from $\{y_j^{GSF}\}_{j=1}^{10}$ as compared with $\{y_j^{unif}\}_{j=1}^{10}$, indicating that delays estimated from data concentrated in the region of information content are more reliable. Model solutions correspond-

Table 1 Estimation of delay τ

	Initial $\hat{\tau}$	$\hat{\tau}$	$SE(\hat{\tau})$
$\{y_j^{unif}\}_{j=1}^{10}$	0.8	0.7862	0.14
$\{y_j^{GSF}\}_{j=1}^{10}$	0.8	1.0247	0.086
$\{y_j^{unif}\}_{j=1}^{10}$	1.2	0.8525	0.13
$\{y_j^{GSF}\}_{j=1}^{10}$	1.2	1.0247	0.086

ing to the estimated $\hat{\tau}$'s overlayed with the data are shown in Fig. 4(a) for simulated uniform data and with data concentrated in [8, 14] in Fig. 4(b).

5.2 Harmonic Oscillator

We turn finally to illustrating the use of the TSF and GSF for the Minorsky harmonic oscillators with delays as given in the Introduction. We recall that the equation with delayed damping has the form

$$\frac{d^2x(t)}{dt^2} + K\frac{dx(t-\tau)}{dt} + bx(t) = g(t), \tag{5.6}$$

while the system with delayed restoring force is given by

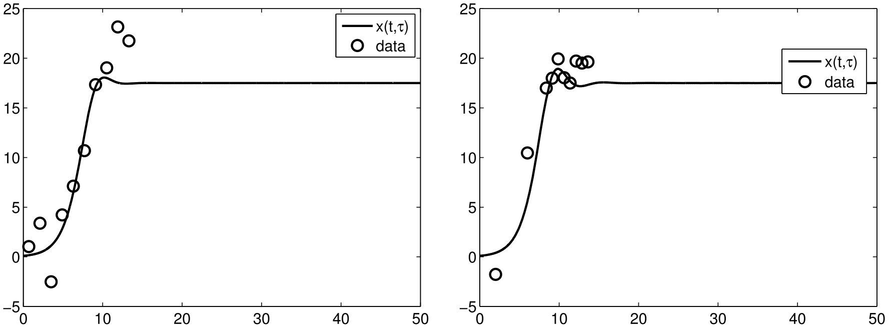
$$\frac{d^2x(t)}{dt^2} + K\frac{dx}{dt} + bx(t-\tau) = g(t). \tag{5.7}$$

We use traditional and generalized sensitivity functions with (5.6) and (5.7) and illustrate their application in determining regions of sensitivity for model parameters K, b and time delay τ. As before, we take the derivative of (5.6) with respect to each parameter q_i, where $q = (K, b, \tau)^T$ to obtain the TSF corresponding to that parameter q_i. First, letting $x = x_1(t)$ and $x_2(t) = \dot{x}(t)$, and rewriting (5.6) as a first order system we have

$$\begin{aligned} \frac{dx_1(t)}{dt} &= x_2(t), \\ \frac{dx_2(t)}{dt} &= g(t) - bx_1(t) - Kx_2(t-\tau). \end{aligned} \tag{5.8}$$

The traditional sensitivity functions are then solutions of

$$\begin{aligned} \frac{ds_1(t)}{dt} &= s_4(t), \\ \frac{ds_2(t)}{dt} &= s_5(t), \end{aligned}$$



(A) model solution: with $\hat{\tau}$ using t_{unif}

(B) model solution: with $\hat{\tau}$ using t_{GSF}

Fig. 4 The solutions to the delay logistic equation with estimated delay $\hat{\tau}$ from data as shown in each graph: (**a**) $\hat{\tau}$ with data corresponding to t_{unif}, (**b**) $\hat{\tau}$ with data corresponding to t_{GSF}

$$\frac{ds_3(t)}{dt} = s_6(t),$$

$$\frac{ds_4(t)}{dt} = -bs_1(t) - Ks_4(t-\tau) - x_2(t-\tau),$$

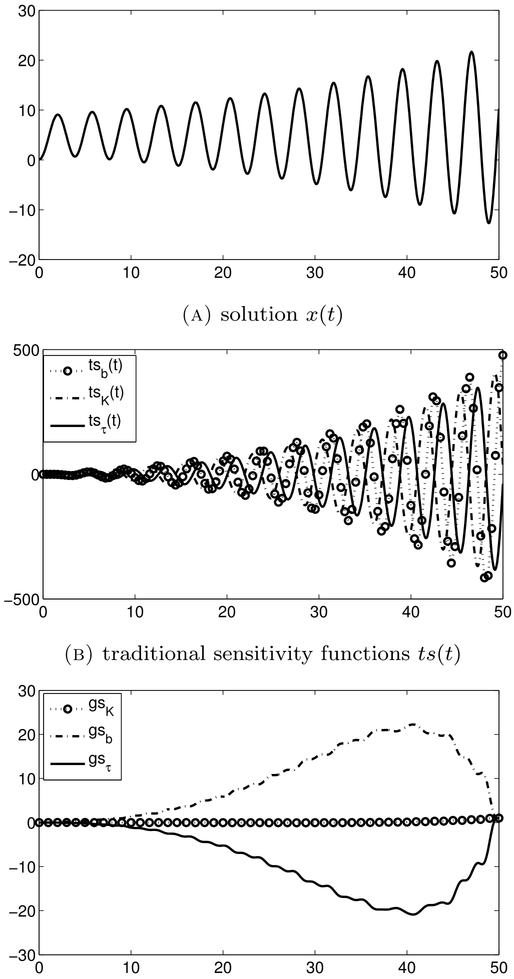
$$\frac{ds_5(t)}{dt} = -bs_2(t) - Ks_5(t-\tau) - x_1(t),$$

$$\frac{ds_6(t)}{dt} = -bs_3(t) - Ks_6(t-\tau) + K\dot{x}_2(t-\tau),$$

for $s_1(t) = \frac{\partial x_1(t)}{\partial K}$, $s_2(t) = \frac{\partial x_1(t)}{\partial b}$, $s_3(t) = \frac{\partial x_1(t)}{\partial \tau}$, $s_4(t) = \frac{\partial x_2(t)}{\partial K}$, $s_5(t) = \frac{\partial x_2(t)}{\partial b}$, and $s_6(t) = \frac{\partial x_2(t)}{\partial \tau}$.

In Fig. 5, the solution for the harmonic oscillator with delayed damping is shown for parameter values $K = 0.5$, $b = 2$, $g(t) \equiv 10$, and delay $\tau = 1$, along with the solutions of the traditional and generalized sensitivity functions with respect to $q = (K, b, \tau)^T$. The solutions of the TSFs imply that the solution is sensitive to all three parameters, growing in magnitude with the amplitude of the oscillations. The sensitivity trajectories indicate the part of the oscillation that is most sensitive to each parameter, indicating which parts are likely due to each force in the harmonic oscillator, and the delay in the damping force. Since the regions of sensitivity overlap, one would be motivated to look at the generalized sensitivity functions to determine whether there are correlation between these three parameters. The solutions to the generalized sensitivity functions clarify this point, and indicate that the parameter b and delay τ are correlated and the parameter K is uncorrelated with the other two over its regions of sensitivity. Therefore, if one were to estimate parameters with this model, one should not expect to estimate both b and τ simultaneously, but estimating either b or τ does not affect one's ability to estimate the parameter K. The solution is not sensitive to any of the parameters until the oscillations grow, indicating that data taken in the beginning time intervals should not be expected to

Fig. 5 Depicted above are (**a**) the solution to the harmonic oscillator with delayed damping $K = 0.5, b = 2, \tau = 1$, and $g(t) = 10$, (**b**) the traditional sensitivity functions and (**c**) the generalized sensitivity functions with respect to K, b, τ



(A) solution $x(t)$

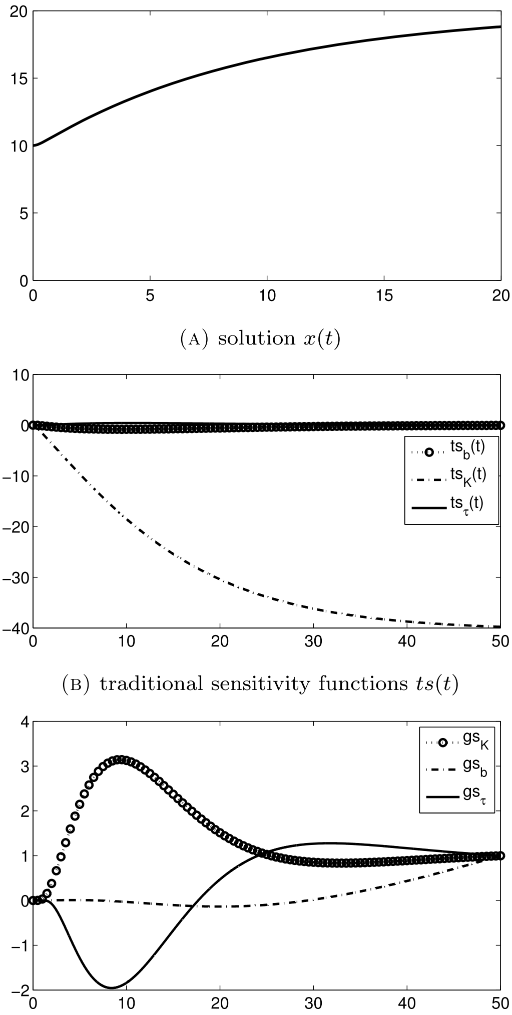
(B) traditional sensitivity functions $ts(t)$

(C) generalized sensitivity functions $gs(t)$

contain much information about any of the parameters. It is not immediately obvious that two parameters K and τ appearing in the same term would be uncorrelated and therefore, both potentially are identifiable from data.

The sensitivity functions for the harmonic oscillator with delayed restoring force, (5.7), are arrived at in the same manner as when the delay appears in the damping term and are therefore omitted. The solution $x(t)$ and the corresponding traditional and generalized sensitivity solutions with respect to $q = (K, b, \tau)^T$ are graphed in Fig. 6. The solution appears to be monotonic, and the traditional sensitivity function solutions indicate that the solution is disproportionately sensitive to the damping coefficient K as compared with b and τ. However, the generalized sensitivity functions provide additional insight in that the solution appears to be sensitive to K early on

Fig. 6 Shown above are (**a**) the solution to the harmonic oscillator with delayed restoring force with $K = 5, b = 0.5, \tau = 1$, and $g(t) = 10$, (**b**) the traditional sensitivity functions and (**c**) the generalized sensitivity functions with respect to K, b, τ



(A) solution $x(t)$

(B) traditional sensitivity functions $ts(t)$

(C) generalized sensitivity functions $gs(t)$

and while there are regions of increased information content relative to the delay τ, that the parameter K is correlated with τ, and one should not expect to identify both simultaneously with data sampled from those intermediate and later regions.

6 Concluding Remarks

After giving a brief survey of previous contributions on theoretical and computational aspects of traditional sensitivity functions for delay differential equation systems, we presented a summary of new theoretical results (proofs for which are given in [28]) for differentiation of solutions with respect to parameters, initial data and delays in general nonlinear delay differential equations. These results provide a theoretical foundation for the rigorous formulation of both traditional and generalized sensitivities for delay systems. We illustrate the ideas in the context of Hutchinson's delayed logistic equation and the classical Minorsky harmonic oscillators with delayed damping or delayed restoring forces.

References

1. J. Arino, L. Wang, G. Wolkowicz, An alternative formulation for a delayed logistic equation. J. Theor. Biol. **241**, 109–118 (2006)
2. C. Baker, F. Rihan, Sensitivity analysis of parameters in modelling with delay-differential equations, *MCCM Tec. Rep.*, **349** (1999), Manchester, ISSN 1360-1725
3. H.T. Banks, *Modeling and Control in the Biomedical Sciences*. Lecture Notes in Biomath., vol. 6 (Springer, Berlin, 1975)
4. H.T. Banks, Delay systems in biological models: approximation techniques, in *Nonlinear Systems and Applications*, ed. by V. Lakshmikantham (Academic Press, New York, 1977), pp. 21–38
5. H.T. Banks, Approximation of nonlinear functional differential equation control systems. J. Optim. Theory Appl. **29**, 383–408 (1979)
6. H.T. Banks, Identification of nonlinear delay systems using spline methods, in *Nonlinear Phenomena in Mathematical Sciences*, ed. by V. Lakshmikantham (Academic Press, New York, 1982), pp. 47–55
7. H.T. Banks, Identification of nonlinear delay systems using spline methods, in *Nonlinear Phenomena in Mathematical Sciences*, ed. by V. Lakshmikanitham (Academic Press, New York, 1982), pp. 47–55
8. H.T. Banks, *A Functional Analysis Framework for Modeling, Estimation and Control in Science and Engineering*, vol. 15 (Taylor & Francis, London, 2012, accepted January 15), 258 pp.
9. H.T. Banks, D.M. Bortz, A parameter sensitivity methodology in the context of HIV delay equation models. J. Math. Biol. **50**(6), 607–625 (2005)
10. H.T. Banks, J.A. Burns, An abstract framework for approximate solutions to optimal control problems governed by hereditary systems, in *International Conference on Differential Equations*, ed. by H. Antosiewicz (Academic Press, San Diego, 1975), pp. 10–25
11. H.T. Banks, J.A. Burns, Hereditary control problems: numerical methods based on averaging approximations. SIAM J. Control Optim. **16**, 169–208 (1978)
12. H.T. Banks, F. Kappel, Spline approximations for functional differential equations. J. Differ. Equ. **34**, 496–522 (1979)
13. H.T. Banks, K. Kunisch, *Estimation Techniques for Distributed Parameter Systems* (Birkhauser, Boston, 1989)
14. H.T. Banks, P.K. Daniel Lamm, Estimation of delays and other parameters in nonlinear functional differential equations. SIAM J. Control Optim. **21**, 895–915 (1983)
15. H.T. Banks, J.M. Mahaffy, Global asymptotic stability of certain models for protein synthesis and repression. Q. Appl. Math. **36**, 209–221 (1978)
16. H.T. Banks, J.M. Mahaffy, Stability of cyclic gene models for systems involving repression. J. Theor. Biol. **74**, 323–334 (1978)
17. H.T. Banks, H. Nguyen, Sensitivity of dynamical systems to Banach space parameters. J. Math. Anal. Appl. **323**, 146–161 (2006). CRSC-TR05-13, February, 2005
18. H.T. Banks, I.G. Rosen, Spline approximations for linear nonautonomous delay systems. J. Math. Anal. Appl. **96**, 226–268 (1983). ICASE Rep. No. 81-33, NASA Langley Res. Center, Oct., 1981

19. H.T. Banks, J.A. Burns, E.M. Cliff, Parameter estimation and identification for systems with delays. SIAM J. Control Optim. **19**, 791–828 (1981)
20. H.T. Banks, S. Dediu, S.L. Ernstberger, Sensitivity functions and their uses in inverse problems. J. Inverse Ill-Posed Probl. **15**, 683–708 (2007). CRSC-TR07-12, July, 2007
21. H.T. Banks, S. Dediu, H.K. Nguyen, Time delay systems with distribution dependent dynamics. Annu. Rev. Control **31**, 17–26 (2007). CRSC-TR06-15, May, 2006
22. H.T. Banks, S. Dediu, H.K. Nguyen, Sensitivity of dynamical systems to parameters in a convex subset of a topological vector space. Math. Biosci. Eng. **4**, 403–430 (2007). CRSC-TR06-25, November, 2006
23. H.T. Banks, M. Davidian, J.R. Samuels Jr., K.L. Sutton, An inverse problem statistical methodology summary, in *Mathematical and Statistical Estimation Approaches in Epidemiology*, ed. by G. Chowell et al. (Springer, Berlin, 2009), pp. 249–302. CRSC-TR08-01, January, 2008, Chap. 11
24. H.T. Banks, J.E. Banks, S.L. Joyner, Estimation in time-delay modeling of insecticide-induced mortality. J. Inverse Ill-Posed Probl. **17**, 101–125 (2009). CRSC-TR08-15, October, 2008
25. H.T. Banks, S. Dediu, S. Ernstberger, F. Kappel, Generalized sensitivities and optimal experimental design. J. Inverse Ill-Posed Probl. **18**, 25–83 (2010). CRSC-TR08-12, revised, November, 2009
26. H.T. Banks, K. Rehm, K. Sutton, Inverse problems for nonlinear delay systems. Methods Appl. Anal. **17**, 331–356 (2010). CRSC-TR10-17, N.C. State University, November, 2010
27. H.T. Banks, K. Holm, F. Kappel, Comparison of optimal design methods in inverse problems. Inverse Probl. **27**, 075002 (2011). CRSC-TR10-11, July, 2010, pp. 31
28. H.T. Banks, D. Robbins, K. Sutton, Theoretical foundations for traditional and generalized sensitivity functions for nonlinear delay differential equations. Math. Biosci. Eng. CRSC-TR12-14, May, 2012 (to appear)
29. V. Barbu, *Nonlinear Semigroups and Differential Equations in Banach Spaces* (Noordhoff, Leyden, 1976)
30. R. Bellman, K.L. Cooke, *Differential-Difference Equations*. Mathematics in Science and Engineering, vol. 6 (Academic Press, New York, 1963)
31. D. Brewer, The differentiability with respect to a parameter of the solution of a linear abstract Cauchy problem. SIAM J. Math. Anal. **13**, 607–620 (1982)
32. D. Brewer, J.A. Burns, E.M. Cliff, Parameter identification for an abstract Cauchy problem by quasilinearization. ICASE Rep. No. 89-75, NASA Langley Res. Center, Oct. 1989
33. J.A. Burns, E.M. Cliff, S.E. Doughty, Sensitivity analysis and parameter estimation for a model of Chlamydia Trachomatis infection. J. Inverse Ill-Posed Probl. **15**, 19–32 (2007)
34. S.N. Busenberg, K.L. Cooke (eds.), *Differential Equations and Applications in Ecology, Epidemics, and Population Problems* (Academic Press, New York, 1981)
35. V. Capasso, E. Grosso, S.L. Paveri-Fontana (eds.), *Mathematics in Biology and Medicine*. Lecture Notes in Biomath., vol. 57 (Springer, Berlin, 1985)
36. J. Caperon, Time lag in population growth response of Isochrysis Galbana to a variable nitrate environment. Ecology **50**, 188–192 (1969)
37. S. Choi, N. Koo, Oscillation theory for delay and neutral differential equations. Trends Math. **2**, 170–176 (1999)
38. K.L. Cooke, Functional differential equations: Some models and perturbation problems, in *Differential Equations and Dynamical Systems*, ed. by J.K. Hale, J.P. LaSalle (Academic Press, New York, 1967), pp. 167–183
39. J.M. Cushing, *Integrodifferential Equations and Delay Models in Population Dynamics*. Lec. Notes in Biomath., vol. 20 (Springer, Berlin, 1977)
40. U. Forys, A. Marciniak-Czochra, Delay logistic equation with diffusion, in *Proc. 8th Nat. Conf. Mathematics Applied to Biology and Medicine*, Lajs, Warsaw, Poland (2002), pp. 37–42
41. U. Forys, A. Marciniak-Czochra, Logistic equations in tumor growth modelling. Int. J. Appl. Math. Comput. Sci. **13**, 317–325 (2003)

42. J.S. Gibson, L.G. Clark, Sensitivity analysis for a class of evolution equations. J. Math. Anal. Appl. **58**, 22–31 (1977)
43. L. Glass, M.C. Mackey, Pathological conditions resulting from instabilities in physiological control systems. Ann. N.Y. Acad. Sci. **316**, 214–235 (1979)
44. K.P. Hadeler, Delay equations in biology, in *Functional Differential Equations and Approximation of Fixed Points*, vol. 730 (Springer, Berlin, 1978), pp. 136–156
45. J.K. Hale, *Theory of Functional Differential Equations* (Springer, New York, 1977)
46. F.C. Hoppensteadt (ed.), *Mathematical Aspects of Physiology*. Lectures in Applied Math, vol. 19 (Am. Math. Soc., Providence, 1981)
47. G.E. Hutchinson, Circular causal systems in ecology. Ann. N.Y. Acad. Sci. **50**, 221–246 (1948)
48. G.E. Hutchinson, *An Introduction to Population Ecology* (Yale University Press, New Haven, 1978)
49. F. Kappel, An approximation scheme for delay equations, in *Nonlinear Phenomena in Mathematical Sciences*, ed. by V. Lakshmikantham (Academic Press, New York, NY, 1982), pp. 585–595
50. F. Kappel, Generalized sensitivity analysis in a delay system. Proc. Appl. Math. Mech. **7**, 1061001–1061002 (2007)
51. F. Kappel, W. Schappacher, Autonomous nonlinear functional differential equations and averaging approximations. Nonlinear Anal. **2**, 391–422 (1978)
52. Y. Kuang, *Delay Differential Equations: With Applications in Population Dynamics* (Academic Press, San Diego, 1993)
53. N. MacDonald, Time lag in a model of a biochemical reaction sequence with end-product inhibition. J. Theor. Biol. **67**, 727–734 (1977)
54. M. Martelli, K.L. Cooke, E. Cumberbatch, B. Tang, H. Thieme (eds.), *Differential Equations and Applications to Biology and to Industry* (World Scientific, Singapore, 1996)
55. J.A.J. Metz, O. Diekmann (eds.), *The Dynamics of Physiologically Structured Populations*. Lecture Notes in Biomath., vol. 68 (Springer, Berlin, 1986)
56. N. Minorsky, Self-excited oscillations in dynamical systems possessing retarded actions. J. Appl. Mech. **9**, A65–A71 (1942)
57. N. Minorsky, On non-linear phenomenon of self-rolling. Proc. Natl. Acad. Sci. **31**, 346–349 (1945)
58. N. Minorsky, *Nonlinear Oscillations* (Van Nostrand, New York, 1962)
59. A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations* (Springer, New York, 1983)
60. D. Robbins, Sensitivity Functions for Delay Differential Equation Models. Ph.D. Dissertation, North Carolina State University, Raleigh, September, 2011
61. K. Schmitt (ed.), *Delay and Functional Differential Equations and Their Applications* (Academic Press, New York, 1972)
62. R. Schuster, H. Schuster, Reconstruction models for the Ehrlich Ascites Tumor of the mouse, in *Mathematical Population Dynamics*, vol. 2, ed. by O. Arino, D. Axelrod, M. Kimmel, Wuertz, Winnipeg (1995), pp. 335–348
63. M.H. Schultz, *Spline Analysis* (Prentice Hall, Englewood Cliffs, 1973)
64. M.H. Schultz, R.S. Varga, L-splines. Numer. Math. **10**, 345–369 (1967)
65. K. Thomseth, C. Cobelli, Generalized sensitivity functions in physiological system identification. Ann. Biomed. Eng. **27**, 606–616 (1999)
66. E.M. Wright, A non-linear difference-differential equation. J. Reine Angew. Math. **494**, 66–87 (1955)

Regularity and Unique Existence of Solution to Linear Diffusion Equation with Multiple Time-Fractional Derivatives

Susanne Beckers and Masahiro Yamamoto

Abstract We consider an initial/boundary value problem for linear diffusion equation with multiple fractional time derivatives and prove the regularity of the solution. The regularity argument implies the unique existence of the solution.

Keywords Fractional diffusion equation · Mild solution · Regularity

Mathematics Subject Classification (2010) 26A33 · 35R10 · 45K05

1 Introduction

Let Ω be a bounded domain in $\mathbb{R}^d$ with sufficiently smooth boundary $\partial\Omega$ and let $0 < \alpha_2 < \alpha_1 < 1$. We consider the following initial/boundary value problem for a diffusion equation with two fractional time derivatives:

$$\partial_t^{\alpha_1} u(x,t) + q(x)\partial_t^{\alpha_2} u(x,t) = (-\mathcal{A}u)(x,t), \quad x \in \Omega,\ t \in (0,T), \tag{1.1}$$

$$u(x,t) = 0, \quad x \in \partial\Omega,\ t \in (0,T) \tag{1.2}$$

and

$$u(x,0) = a(x), \quad x \in \Omega. \tag{1.3}$$

The first named author thanks JASSO Honors Scholarship and the Global COE Program "The Research and Training Center for New Development in Mathematics" at Graduate School of Mathematical Sciences of The University of Tokyo for financial support.

S. Beckers
Faculty of Mathematics, Lothar Collatz School for Computing in Science, Universität Hamburg, Gindelberg 5, 20144 Hamburg, Germany
e-mail: susnne.beckers@uni-hamburg.de

M. Yamamoto (✉)
Graduate School of Mathematical Sciences, The University of Tokyo, 3-8-9 Komaba Meguro, Tokyo 153-8914, Japan
e-mail: myama@ms.u-tokyo.ac.jp

K. Bredies et al. (eds.), *Control and Optimization with PDE Constraints*,
International Series of Numerical Mathematics 164,
DOI 10.1007/978-3-0348-0631-2_3, © Springer Basel 2013

Here, for $0 < \alpha < 1$, we denote by ∂_t^α the Caputo fractional derivative with respect to t:

$$\partial_t^\alpha g(t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t (t-\tau)^{-\alpha} \frac{d}{d\tau} g(\tau) d\tau$$

where Γ is the Gamma function and $q \in W^{2,\infty}(\Omega)$. The space $W^{2,\infty}(\Omega)$ is the usual Sobolev space (Adams [1]). Moreover the operator $-\mathcal{A}$ is a symmetric uniformly elliptic operator, that is,

$$(-\mathcal{A}u)(x) = \sum_{i=1}^d \frac{\partial}{\partial x_i}\left(\sum_{j=1}^d a_{ij}(x)\frac{\partial}{\partial x_j}u(x)\right) + b(x)u(x), \quad x \in \Omega,$$

where $a_{ij} = a_{ji}$, $1 \le i, j \le d$, $a_{ij} \in C^1(\overline{\Omega})$, $b \in C(\overline{\Omega})$, $b(x) \le 0$ for $x \in \overline{\Omega}$, and we assume that there exists a constant $C_0 > 0$ such that

$$C_0 \sum_{i=1}^d \xi_i^2 \le \sum_{i,j=1}^d A_{ij}(x)\xi_i\xi_j, \quad x \in \overline{\Omega},\ \xi \in \mathbb{R}^d.$$

In the special case $q \equiv 0$, (1.1) is a diffusion equation with a single fractional time derivative. For such equations there exists a large and rapidly growing number of publications which we do not intend to list completely: Bazhlekova [3, 4], Eidelman and Kochubei [6], Luchko [12, 13], Prüss [21], Sakamoto and Yamamoto [22]. Also see Agarwal [2], Fujita [7], Gejji and Jafari [8], Mainardi [15–17], Nigmatullin [18], Schneider and Wyss [23].

In this article, we consider the case of multiple fractional time derivatives. Such equations can be considered as more feasible model equations than equations with a single fractional time derivative in modeling diffusion in porous media. In the case where the functions in (1.1) are not dependent on x, we refer to several works and refer for example to Diethelm and Luchko [5], Podlubny [20], Chap. 3, for instance. In particular, in [5], some physical interpretations are given. As for diffusion equations with multiple fractional time derivatives, see Jiang, Liu, Turner and Burrage [10] and Luchko [14] which argue a general number of time fractional derivatives. The article[10] discusses the spatially one dimensional case with constant coefficients where also the spatial fractional derivative is considered, and establishes the formula of the solution, and [14] assumes that the coefficients of the time derivatives are constant to prove unique existence of the solution by the Fourier method as well as the maximum principle and related properties.

Unlike [10] and [14], we treat a general case of x-dependent coefficient of fractional time derivatives. In our case, we cannot apply the Fourier method or obtain an analytic solution. We apply the perturbation method and the theory of evolution equations to prove regularity as well as unique existence of solution to (1.1)–(1.3). Such results should be the starting point for further research concerning the theory of nonlinear fractional diffusion equations, numerical analysis, control theory and inverse problems. In forthcoming papers, we will discuss those subjects.

This paper is composed of four sections including the current section. In Sect. 2, we present the main result for the case of two fractional time derivatives and in Sect. 3 we prove it. Section 4 is devoted to the case of general multiple fractional time derivatives.

2 Main Results

Let $L^2(\Omega)$ be the usual L^2-space with the scalar product $(\cdot,\cdot)$, and $H^\ell(\Omega)$, $H_0^m(\Omega)$ denote the usual Sobolev spaces (e.g., Adams [1]). We set $\|a\|_{L^2(\Omega)} = (a,a)^{\frac{1}{2}}$.

We define the operator A in $L^2(\Omega)$ by

$$(Au)(x) = (\mathcal{A}u)(x), \quad x \in \Omega, \qquad \mathcal{D}(A) = H^2(\Omega) \cap H_0^1(\Omega).$$

Then the fractional power A^γ is defined for $\gamma \in \mathbb{R}$ (see for instance [19]), and $\mathcal{D}(A^\gamma) \subset H^{2\gamma}(\Omega)$, $\mathcal{D}(A^{\frac{1}{2}}) = H_0^1(\Omega)$ for example. We note that $\|u\|_{\mathcal{D}(A^\gamma)} := \|A^\gamma u\|_{L^2(\Omega)}$ is a stronger norm than $\|u\|_{L^2(\Omega)}$ for $\gamma > 0$.

Since $-A$ is a symmetric uniformly elliptic operator, the spectrum of A is entirely composed of eigenvalues and counting according to the multiplicities, we can set: $0 < \lambda_1 \le \lambda_2 \le \cdots$. By $\phi_n \in H^2(\Omega) \cap H_0^1(\Omega)$, we denote the orthonormal eigenfunction corresponding to λ_n: $A\phi_n = \lambda_n \phi_n$. Then the sequence $\{\phi_n\}_{n\in\mathbb{N}}$ is an orthonormal basis in $L^2(\Omega)$. Moreover, we see that

$$\mathcal{D}\big(A^\gamma\big) = \left\{ \psi \in L^2(\Omega); \sum_{n=1}^{\infty} \lambda_n^{2\gamma} \big|(\psi,\phi_n)\big|^2 < \infty \right\}$$

and that $\mathcal{D}(A^\gamma)$ is a Hilbert space with the norm

$$\|\psi\|_{\mathcal{D}(A^\gamma)} = \left\{ \sum_{n=1}^{\infty} \lambda_n^{2\gamma} \big|(\psi,\phi_n)\big|^2 \right\}^{\frac{1}{2}}.$$

Henceforth we associate with $u(x,t)$, provided that it is well-defined, a map $u(\cdot) : (0,T) \longrightarrow L^2(\Omega)$ by $u(t)(x) = u(x,t)$, $0 < t < T$, $x \in \Omega$. Then we can write (1.1)–(1.3) as

$$\partial_t^{\alpha_1} u(t) + q\partial_t^{\alpha_2} u(t) = -Au(t), \quad t > 0 \text{ in } L^2(\Omega), \tag{2.1}$$

$$u(0) = a \in L^2(\Omega).$$

Remark 1 The interpretation of the initial condition should be made in a suitable function space. In our case, as Theorem 1 asserts, we have $\lim_{t\to 0} \|u(t) - a\|_{L^2(\Omega)} = 0$.

Moreover we define the Mittag–Leffler function $E_{\alpha,\beta}$ by

$$E_{\alpha,\beta}(z) := \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k+\beta)}, \quad z \in \mathbb{C},$$

where $\alpha > 0$ and $\beta \in \mathbb{R}$ are arbitrary constants. Using the power series, we can directly verify that $E_{\alpha,\beta}(z)$ is an entire function.

Now we define the operators $S(t) : L^2(\Omega) \to L^2(\Omega)$, $t \geq 0$, by

$$S(t)a := \sum_{n=1}^{\infty} (a, \phi_n) E_{\alpha_1,1}\big(-\lambda_n t^{\alpha_1}\big)\phi_n \quad \text{in } L^2(\Omega) \tag{2.2}$$

for $a \in L^2(\Omega)$. Then we can prove that $S(t) : L^2(\Omega) \longrightarrow L^2(\Omega)$ is a bounded linear operator for $t \geq 0$ (e.g., Sakamoto and Yamamoto [22]). Moreover, termwise differentiation is possible and gives

$$S'(t)a = -\sum_{n=0}^{\infty} \lambda_n (a, \phi_n) t^{\alpha_1-1} E_{\alpha_1,\alpha_1}\big(-\lambda_n t^{\alpha_1}\big)\phi_n \quad \text{in } L^2(\Omega) \tag{2.3}$$

and

$$S''(t)a = -\sum_{n=0}^{\infty} \lambda_n (a, \phi_n) t^{\alpha_1-2} E_{\alpha_1,\alpha_1-1}\big(-\lambda_n t^{\alpha_1}\big)\phi_n \quad \text{in } L^2(\Omega) \tag{2.4}$$

for $a \in L^2(\Omega)$.

For $F \in L^2(\Omega \times (0,T))$ and $a \in L^2(\Omega)$, there exists a unique solution in a suitable class (e.g., Sakamoto and Yamamoto [22]) to the problem

$$\partial_t^{\alpha_1} u(t) = -Au(t) + F, \quad 0 < t < T, \tag{2.5}$$

$$u(0) = a. \tag{2.6}$$

This solution is given by

$$u(t) = \int_0^t A^{-1} S'(t-\tau) F(\tau) d\tau + S(t)a, \quad t > 0. \tag{2.7}$$

In view of (2.7), we mainly discuss the equation

$$u(t) = S(t)a - \int_0^t A^{-1} S'(t-\tau) q \partial_t^{\alpha_2} u(\tau) d\tau, \quad 0 < t < T, \tag{2.8}$$

in order to establish unique existence of solutions to (2.1). Henceforth C denotes generic positive constants which are independent of a in (1.2), but may depend on T, α_1, α_2 and the coefficients of the operator A and q.

We can state our first main result.

Theorem 1 *We assume that* $u \in C((0,T]; L^2(\Omega))$ *satisfies* (2.8) *and*

$$\alpha_1 + \alpha_2 > 1.$$

Then

$$\|u(t)\|_{H^{2\gamma}(\Omega)} \le Ct^{-\alpha_1\gamma}\|a\|_{L^2(\Omega)}, \quad 0 < t \le T$$

for any $\gamma \in (0,1)$.

We may be able to remove the condition $\alpha_1 + \alpha_2 > 1$. On the other hand, Prüss established regularity in case $\gamma = 1$ for general $\alpha_1, \alpha_2 \in (0,1)$ under a strong condition on $a \in \mathcal{D}(A)$ (see [21], in particular the perturbation theorem on p. 60).

On the basis of Theorem 1, a standard argument (e.g., Henry [9]) yields

Theorem 2 *For any* $\gamma \in (0,1)$ *there exists a mild solution to* (2.8) *in the space* $C((0,T]; \mathcal{D}(A^\gamma)) \cap C([0,T]; L^2(\Omega))$.

3 Proof of Theorem 1

First we have

$$A^{\gamma-1}S'(t)a = -t^{\alpha_1-1}\sum_{n=1}^{\infty}\lambda_n^\gamma(a,\phi_n)E_{\alpha_1,\alpha_1}\big(-\lambda_n t^{\alpha_1}\big)\phi_n \quad \text{in } L^2(\Omega) \tag{3.1}$$

for $a \in L^2(\Omega)$ and $\gamma \ge 0$. Moreover, since

$$\big|E_{\alpha_1,\alpha_1}(-\eta)\big| \le \frac{C}{1+\eta}, \quad \eta > 0$$

(e.g., Theorem 1.6 on p. 35 in Podlubny [20]), we can prove

$$\big\|A^{\gamma-1}S'(t)\big\| \le Ct^{\alpha_1-1-\alpha_1\gamma}, \quad t > 0 \tag{3.2}$$

and

$$\big\|A^{-1}S''(t)\big\| \le Ct^{\alpha_1-2}, \quad t > 0. \tag{3.3}$$

Now we proceed to the proof of Theorem 1. We set

$$v(t) := \int_0^t A^{\gamma-1}S'(t-\eta)q\partial_t^{\alpha_2}u(\eta)d\eta, \quad 0 < t < T.$$

By (2.8), we have

$$A^{\gamma} u(t) = A^{\gamma} S(t)a - v(t), \quad 0 < t < T.$$

Therefore, using

$$\|u(t)\|_{H^{2\gamma}(\Omega)} \le C \|A^{\gamma} u(t)\|_{L^2(\Omega)}$$

(see the beginning of Sect. 2), it is sufficient to estimate $\|A^{\gamma} S(t)a\|_{L^2(\Omega)} + \|v(t)\|_{L^2(\Omega)}$. First we will estimate $\|v(t)\|_{L^2(\Omega)}$. Substituting the definition of $\partial_t^{\alpha_2} u$ and changing the order of integration, we have

$$\begin{aligned} v(t) &= \int_0^t A^{\gamma-1} S'(t-\eta) \frac{1}{\Gamma(1-\alpha_2)} \left(\int_0^{\eta} (\eta-\tau)^{-\alpha_2} q u'(\tau) d\tau \right) d\eta \\ &= \frac{1}{\Gamma(1-\alpha_2)} \int_0^t H(t,\tau) q u'(\tau) d\tau, \quad 0 < t < T. \end{aligned} \tag{3.4}$$

Here we have set

$$H(t,\tau) = \int_{\tau}^{t} A^{\gamma-1} S'(t-\eta)(\eta-\tau)^{-\alpha_2} d\eta.$$

Decomposing the integrand and introducing the change of variables $\eta - \tau \to \eta$ we obtain

$$\begin{aligned} H(t,\tau) &= \int_{\tau}^{t} A^{\gamma-1} S'(t-\eta)(\eta-\tau)^{-\alpha_2} d\eta \\ &= \int_{\tau}^{t} A^{\gamma-1} S'(t-\eta)\big[(\eta-\tau)^{-\alpha_2} - (t-\tau)^{-\alpha_2}\big] d\eta \\ &\quad + \int_{\tau}^{t} A^{\gamma-1} S'(t-\eta) d\eta (t-\tau)^{-\alpha_2} \\ &= \int_0^{t-\tau} A^{\gamma-1} S'(t-\eta-\tau)\big[\eta^{-\alpha_2} - (t-\tau)^{-\alpha_2}\big] d\eta \\ &\quad + \int_{\tau}^{t} A^{\gamma-1} S'(t-\eta) d\eta (t-\tau)^{-\alpha_2} \\ &= \int_0^{t-\tau} A^{\gamma-1} S'(t-\eta-\tau)\big[\eta^{-\alpha_2} - (t-\tau)^{-\alpha_2}\big] d\eta \\ &\quad + A^{\gamma-1} S(0)(t-\tau)^{-\alpha_2} - A^{\gamma-1} S(t-\tau)(t-\tau)^{-\alpha_2} \\ &:= I_1(t,\tau) + I_2(t,\tau). \end{aligned} \tag{3.5}$$

On the other hand, we have

$$\begin{aligned}\partial_\tau I_1(t,\tau) = &-\int_0^{t-\tau} A^{\gamma-1}S''(t-\eta-\tau)\big(\eta^{-\alpha_2}-(t-\tau)^{-\alpha_2}\big)d\eta \\ &-\alpha_2\int_0^{t-\tau} A^{\gamma-1}S'(t-\eta-\tau)(t-\tau)^{-\alpha_2-1}d\eta \\ &-\lim_{\eta\to t-\tau} A^{\gamma-1}S'(t-\tau-\eta)\big[\eta^{-\alpha_2}-(t-\tau)^{-\alpha_2}\big].\end{aligned}$$

By the estimate (3.2) we obtain

$$\begin{aligned}&\big\|A^{\gamma-1}S'(t-\tau-\eta)\big(\eta^{-\alpha_2}-(t-\tau)^{-\alpha_2}\big)\big\|_{L^2(\Omega)} \\ &\le C(t-\tau-\eta)^{\alpha_1-1-\alpha_1\gamma}\frac{|(t-\tau)^{\alpha_2}-\eta^{\alpha_2}|}{\eta^{\alpha_2}(t-\tau)^{\alpha_2}}.\end{aligned}$$

According to the mean value theorem, we can choose $\theta\in(\eta,t-\tau)$ such that

$$\big|(t-\tau)^{\alpha_2}-\eta^{\alpha_2}\big| = \big|\alpha_2\theta^{\alpha_2-1}(t-\tau-\eta)\big| \le \alpha_2\eta^{\alpha_2-1}(t-\tau-\eta).$$

Hence we obtain

$$\begin{aligned}&\big\|A^{\gamma-1}S'(t-\tau-\eta)\big(\eta^{-\alpha_2}-(t-\tau)^{-\alpha_2}\big)\big\|_{L^2(\Omega)} \\ &\le C\alpha_2\eta^{-1}(t-\tau)^{-\alpha_2}(t-\tau-\eta)^{\alpha_1-\alpha_1\gamma}\longrightarrow 0 \quad \text{as } \eta\to t-\tau\end{aligned}$$

by $\alpha_1-\alpha_1\gamma>0$. This implies

$$\begin{aligned}\partial_\tau I_1(t,\tau) = &-\int_0^{t-\tau} A^{\gamma-1}S''(t-\eta-\tau)\big(\eta^{-\alpha_2}-(t-\tau)^{-\alpha_2}\big)d\eta \\ &-\alpha_2\int_0^{t-\tau} A^{\gamma-1}S'(t-\eta-\tau)(t-\tau)^{-\alpha_2-1}d\eta, \quad 0<t<T. \end{aligned} \tag{3.6}$$

On the other hand, we have

$$\begin{aligned}\partial_\tau I_2(t,\tau) &= -\alpha_2 A^{\gamma-1}S(t-\tau)(t-\tau)^{-\alpha_2-1} + A^{\gamma-1}S(0)\alpha_2(t-\tau)^{-\alpha_2-1} \\ &\quad + A^{\gamma-1}S'(t-\tau)(t-\tau)^{-\alpha_2} \\ &= \alpha_2\int_0^{t-\tau} A^{\gamma-1}S'(t-\eta-\tau)(t-\tau)^{-\alpha_2-1}d\eta \\ &\quad + A^{\gamma-1}S'(t-\tau)(t-\tau)^{-\alpha_2}.\end{aligned}$$

Adding this and (3.6) we obtain

$$\begin{aligned}\partial_\tau H(t,\tau) = &-\int_0^{t-\tau} A^{\gamma-1}S''(t-\eta-\tau)\big(\eta^{-\alpha_2}-(t-\tau)^{-\alpha_2}\big)d\eta \\ &+ A^{\gamma-1}S'(t-\tau)(t-\tau)^{-\alpha_2}.\end{aligned} \tag{3.7}$$

Using (3.7) in (3.4), integrating by parts and using $H(t,t)=0$ we obtain

$$\begin{aligned}\big(\Gamma(1-\alpha_2)\big)v(t) &= \int_0^t H(t,\tau)qu'(\tau)d\tau\\ &= -H(t,0)qa\\ &\quad + \int_0^t \bigg[\int_0^{t-\tau} A^{\gamma-1}S''(t-\eta-\tau)\big(\eta^{-\alpha_2}-(t-\tau)^{-\alpha_2}\big)d\eta\\ &\quad - A^{\gamma-1}S'(t-\tau)(t-\tau)^{-\alpha_2}\bigg]qu(\tau)d\tau\\ &:= I_3(t)+I_4(t).\end{aligned}$$

We set

$$B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},\quad \alpha,\beta>0.$$

First, by (3.2) and $q\in W^{2,\infty}(\Omega)$ we have

$$\begin{aligned}\|I_3(t)\|_{L^2(\Omega)} = \|-H(t,0)qa\|_{L^2(\Omega)} &= \bigg\|-\int_0^t A^{\gamma-1}S'(t-\eta)\eta^{-\alpha_2}d\eta qa\bigg\|_{L^2(\Omega)}\\ &\le C\|a\|_{L^2(\Omega)}\int_0^t (t-\eta)^{\alpha_1-\alpha_1\gamma-1}\eta^{-\alpha_2}d\eta\\ &= C\|a\|_{L^2(\Omega)}B(1-\alpha_2,\alpha_1-\alpha_1\gamma)t^{\alpha_1-\alpha_1\gamma-\alpha_2}, \qquad (3.8)\end{aligned}$$

since $1-\alpha_2>0$ and $\alpha_1-\alpha_1\gamma>0$.

On the other hand, by $q\in W^{2,\infty}(\Omega)$ and $u|_{\partial\Omega}=0$, we have

$$\|A\big(qu(\tau)\big)\|_{L^2(\Omega)} \le C\|qu(\tau)\|_{H^2(\Omega)} \le C\|u(\tau)\|_{H^2(\Omega)} \le C\|Au(\tau)\|_{L^2(\Omega)}$$

and $\|qu(\tau)\|_{L^2(\Omega)}\le C\|u(\tau)\|_{L^2(\Omega)}$, that is,

$$\|A^0\big(qu(\tau)\big)\|_{L^2(\Omega)} \le C\|A^0u(\tau)\|_{L^2(\Omega)}.$$

Hence the interpolation theorem (see for instance Lions and Magenes [11], Theorem 5.1 on p. 27) we obtain

$$\|A^\gamma\big(qu(\tau)\big)\|_{L^2(\Omega)} \le C\|A^\gamma u(\tau)\|_{L^2(\Omega)}.$$

Therefore by (3.2) and (3.3), the second term of $I_4(t)$ can be estimated as follows:

$$\begin{aligned}\|I_4(t)\|_{L^2(\Omega)} \le C\int_0^t \bigg[&\int_0^{t-\tau}(t-\eta-\tau)^{\alpha_1-2}\big(\eta^{-\alpha_2}-(t-\tau)^{-\alpha_2}\big)d\eta\\ &+(t-\tau)^{\alpha_1-1-\alpha_2}\bigg]\|A^\gamma\big(qu(\tau)\big)\|_{L^2(\Omega)}d\tau\end{aligned}$$

$$
\begin{aligned}
&\le C\int_0^t\Bigg[\int_0^{t-\tau}(t-\eta-\tau)^{\alpha_1-2}\frac{(t-\tau-\eta)^{\alpha_2}}{\eta^{\alpha_2}(t-\tau)^{\alpha_2}}d\eta\\
&\quad+(t-\tau)^{\alpha_1-1-\alpha_2}\Bigg]\big\|A^\gamma u(\tau)\big\|_{L^2(\Omega)}d\tau\\
&\le C\int_0^t\Bigg[\int_0^{t-\tau}(t-\eta-\tau)^{\alpha_1+\alpha_2-2}\eta^{-\alpha_2}d\eta\\
&\quad+(t-\tau)^{\alpha_1-1-\alpha_2}\Bigg]\big\|A^\gamma u(\tau)\big\|_{L^2(\Omega)}d\tau\\
&=C\int_0^t\big(B(1-\alpha_2,\alpha_1+\alpha_2-1)(t-\tau)^{\alpha_1-1}\\
&\quad+(t-\tau)^{\alpha_1-1-\alpha_2}\big)\big\|A^\gamma u(\tau)\big\|_{L^2(\Omega)}d\tau.
\end{aligned}
$$

For the last equality, we used $\alpha_1+\alpha_2>1$. Therefore we have

$$
\begin{aligned}
\big\|\Gamma(1-\alpha_2)v(t)\big\|_{L^2(\Omega)}\le{}& C\|a\|_{L^2(\Omega)}^2B(1-\alpha_2,\alpha_1-\alpha_1\gamma)t^{\alpha_1-\alpha_1\gamma-\alpha_2}\\
&+C\int_0^t(t-\tau)^{\alpha_1-1-\alpha_2}\big\|A^\gamma u(\tau)\big\|_{L^2(\Omega)}d\tau.
\end{aligned}
$$

Thus the estimate of $\|v(t)\|_{L^2(\Omega)}$ is completed.

Next we estimate $\|A^\gamma S(t)a\|_{L^2(\Omega)}$. By Theorem 1.6 (p. 35) in [20], we obtain

$$
\begin{aligned}
\big\|A^\gamma S(t)a\big\|_{L^2(\Omega)}^2&=\left\|\sum_{n=1}^\infty(a,\phi_n)\lambda_n^\gamma E_{\alpha_1,1}\big(-\lambda_nt^{\alpha_1}\big)\phi_n\right\|_{L^2(\Omega)}^2\\
&\le C\sum_{n=1}^\infty(a,\phi_n)^2t^{-2\alpha_1\gamma}\left(\frac{(\lambda_nt^{\alpha_1})^\gamma}{1+\lambda_nt^{\alpha_1}}\right)^2\\
&\le Ct^{-2\alpha_1\gamma}\|a\|_{L^2(\Omega)}^2,
\end{aligned}
$$

and hence

$$
\begin{aligned}
\big\|A^\gamma u(t)\big\|_{L^2(\Omega)}\le{}& C\|a\|_{L^2(\Omega)}\big(t^{-\alpha_1\gamma}+t^{\alpha_1-\alpha_1\gamma-\alpha_2}\big)\\
&+C\int_0^t(t-\tau)^{\alpha_1-1-\alpha_2}\big\|A^\gamma u(\tau)\big\|_{L^2(\Omega)}d\tau\\
\le{}& C\|a\|_{L^2(\Omega)}t^{-\alpha_1\gamma}+C\int_0^t(t-\tau)^{\alpha_1-1-\alpha_2}\big\|A^\gamma u(\tau)\big\|_{L^2(\Omega)}d\tau,\\
&0<t<T.
\end{aligned}
$$

Therefore by an inequality of Gronwall type (see [9], Exercise 3 (p. 190)), we obtain

$$
\big\|A^\gamma u(t)\big\|_{L^2(\Omega)}\le C\|a\|_{L^2(\Omega)}t^{-\alpha_1\gamma},\quad 0<t\le T.
$$

Thus the proof is completed.

4 Generalization

The results in Sect. 3 shall now be extended to the solution of linear diffusion equation with multiple fractional time derivatives:

$$\partial_t^{\alpha_1} u(t) + \sum_{j=2}^{\ell} q_j \partial_t^{\alpha_j} u(t) = -Au(t), \quad t > 0$$

and

$$u(0) = a \in L^2(\Omega),$$

where $0 < \alpha_\ell < \cdots < \alpha_2 < \alpha_1 < 1$ and $q_j \in W^{2,\infty}(\Omega)$, $2 \le j \le \ell$.

As before the lower-order derivatives are regarded as source terms and we consider

$$u(t) = S(t)a - \int_0^t A^{-1} S'(t-\tau) \sum_{j=2}^{\ell} q_j \partial_t^{\alpha_j} u(\tau) d\tau, \quad 0 < t < T. \tag{4.1}$$

Similarly to Theorem 1, we can prove

Theorem 3 *We assume that $u \in C((0,T]; L^2(\Omega))$ satisfies (4.1) and*

$$0 < \alpha_\ell < \cdots < \alpha_1, \quad \alpha_1 + \alpha_\ell > 1.$$

Then

$$\left\| u(t) \right\|_{H^{2\gamma}(\Omega)} \le C t^{-\alpha_1 \gamma} \|a\|_{L^2(\Omega)}, \quad 0 < t \le T$$

for any $\gamma \in (0,1)$. Moreover there exists a mild solution to (4.1) in the space $C((0,T]; \mathcal{D}(A^\gamma)) \cap C([0,T]; L^2(\Omega))$ with $\gamma \in (0,1)$.

Acknowledgements The authors thank Mr. Zhiyuan Li (The University of Tokyo) for his careful reading and remarks, and the anonymous referees for valuable comments which are useful for improving the manuscript.

References

1. R.A. Adams, *Sobolev Spaces* (Academic Press, New York, 1975)
2. O.P. Agarwal, Solution for a fractional diffusion-wave equation defined in a bounded domain. Nonlinear Dyn. **29**, 145–155 (2002)
3. E. Bazhlekova, The abstract Cauchy problem for the fractional evolution equation. Fract. Calc. Appl. Anal. **1**, 255–270 (1998)
4. E. Bazhlekova, Fractional Evolution Equation in Banach Spaces, Doctoral Thesis, Eindhoven University of Technology, 2001
5. K. Diethelm, Y. Luchko, Numerical solution of linear multi-term initial value problems of fractional order. J. Comput. Anal. Appl. **6**, 243–263 (2004)

6. S.D. Eidelman, A.N. Kochubei, Cauchy problem for fractional diffusion equations. J. Differ. Equ. **199**, 211–255 (2004)
7. Y. Fujita, Integrodifferential equation which interpolates the heat equation and the wave equation. Osaka J. Math. **27**, 309–321, 797–804 (1990)
8. V.D. Gejji, H. Jafari, Boundary value problems for fractional diffusion-wave equation. Aust. J. Math. Anal. Appl. **3**, 1–8 (2006)
9. D. Henry, *Geometric Theory of Semilinear Parabolic Equations* (Springer, Berlin, 1981)
10. H. Jiang, F. Liu, I. Turner, K. Burrage, Analytical solutions for the multi-term time-space Caputo-Riesz fractional advection-diffusion equations on a finite domain. J. Math. Anal. Appl. **389**, 1117–1127 (2012)
11. J.L. Lions, E. Magenes, *Non-homogeneous Boundary Value Problems and Applications* (Springer, Berlin, 1972)
12. Y. Luchko, Maximum principle for the generalized time-fractional diffusion equation. J. Math. Anal. Appl. **351**, 218–223 (2009)
13. Y. Luchko, Some uniqueness and existence results for the initial-boundary-value problems for the generalized time-fractional diffusion equation. Comput. Math. Appl. **59**, 1766–1772 (2010)
14. Y. Luchko, Initial-boundary-value problems for the generalized multi-term time-fractional diffusion equation. J. Math. Anal. Appl. **374**, 538–548 (2011)
15. F. Mainardi, On the initial value problem for the fractional diffusion-wave equation, in *Waves and Stability in Continuous Media*, ed. by S. Rionero, T. Ruggeri (World Scientific, Singapore, 1994), pp. 246–251
16. F. Mainardi, The time fractional diffusion-wave equation. Radiophys. Quantum Electron. **38**, 13–24 (1995)
17. F. Mainardi, The fundamental solutions for the fractional diffusion-wave equation. Appl. Math. Lett. **9**, 23–28 (1996)
18. R.R. Nigmatullin, The realization of the generalized transfer equation in a medium with fractal geometry. Phys. Status Solidi B **133**, 425–430 (1986)
19. A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations* (Springer, Berlin, 1983)
20. I. Podlubny, *Fractional Differential Equations* (Academic Press, San Diego, 1999)
21. J. Prüss, *Evolutionary Integral Equations and Applications* (Birkhäuser, Basel, 1993)
22. K. Sakamoto, M. Yamamoto, Initial value/boundary value problems for fractional diffusion-wave equations and applications to some inverse problems. J. Math. Anal. Appl. **382**, 426–447 (2011)
23. W.R. Schneider, W. Wyss, Fractional diffusion and wave equations. J. Math. Phys. **30**, 134–144 (1989)

Nonsmooth Optimization Method and Sparsity

Kazufumi Ito

Abstract Nonsmooth variational problems are analyzed using the Lagrange multiplier theory. In particular, the sparsity optimization method has a multitude of important applications, i.e., in imaging analysis and friction contact and inverse problems, and can be cast as nonsmooth variational problems. The optimality condition is derived and it is of the form of the complementarity systems. An effective numerical optimization method using the semismooth Newton method is then developed and analyzed. The method takes the form of primal-dual active set methods and is much more efficient than numerical optimization algorithms based on first order methods. The ℓ^0 sparsity optimization for the linear least square problem is considered. The necessary optimality condition is derived and a numerical algorithm based on the Lagrange multiplier rule to determine a solution is developed and analyzed.

Keywords Nonsmooth optimization · Sparsity optimization · Mixed integer programming · Primal-dual active set method

1 Introduction

The class of variational problems we investigate can be written as

$$\min \quad F(y)+\varphi(\Lambda y) \quad \text{over } y\in\mathcal{C}, \tag{1.1}$$

where X, H are real Hilbert spaces, $\mathcal{C}$ is a closed convex subset of X, and $\Lambda\in\mathcal{L}(X,H)$. We identify H with its dual space. Furthermore, $F:X\to R^+$ is a continuously differentiable, convex function, and $\varphi:H\to(-\infty,\infty]$ is a proper, lower semi-continuous, convex but not necessarily differentiable function. This problem class encompasses a wide variety of optimization problems. An important class

The author is partially supported by US-ARO grant 49308-MA, and US-AFSOR grant FA9550-06-1-0241.

K. Ito (✉)
Department of Mathematics and Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC, USA
e-mail: kito@math.ncsu.edu

K. Bredies et al. (eds.), *Control and Optimization with PDE Constraints*,
International Series of Numerical Mathematics 164,
DOI 10.1007/978-3-0348-0631-2_4, © Springer Basel 2013

of concrete problems that can be formulated as (1.1) is discussed in Sect. 2. The Lagrange multiplier approach and the optimality system for analyzing (1.1) is described in Sect. 3. A semismooth Newton method applied to the optimality system is developed and analyzed for the bilateral inequality constraint and L^1-type optimization in Sect. 4.

There is an increasing literature of applications of sparsity optimization, e.g., [2, 4, 5, 22] and the references therein, and we refer to [2, 21, 22] for the mathematical analysis of sparsity optimization. For $A \in \mathcal{L}(\ell^2, \ell^2)$, $b \in \ell^2$ and $\beta > 0$ we consider the minimization of

$$\frac{1}{2}|Ax-b|^2+\beta N_0(x), \tag{1.2}$$

where for $x \in \ell^2$

$$N_0(x)=\sum_k |x_k|^0$$

is the counting measure of a number sequence $x \in \ell^2$, i.e., $0^0=0$ and $s^0=1$ if $s>0$. The sparsity method provides an efficient way to extract essential components of solutions and is used for data compression and order reduction methods in a wide class of applications, including signal and image analysis, inverse scattering, deconvolution and tomography problems, and wavelet and generalized Fourier analysis. The case when the ℓ^1 norm is replaced by $N_0(x)$ in (1.2) is the most standard and well-analyzed, and it can be analyzed by the convex formulation (1.1). But, since N_0 is not convex, our analysis and method for (1.1) cannot be applied directly.

For a data mining problems, i.e., problem of determining the most reliable measurements ($k \in K$, where K is a set of indexes) from data $b \in \ell_2$ for the reconstruction of u by $(Au-b)_{k\in K}=0$, we formulate the minimization of the form

$$N_0(Au-b)+\frac{\alpha}{2}|u|_2^2. \tag{1.3}$$

In Sect. 5 we derive the complementarity system for the optimality condition and the Lagrange formulation and develop the primal-dual active set method.

2 Applications and Concrete Examples

In this section we discuss examples of nonsmooth optimization problems and potential applications.

2.1 Variational Inequalities

Consider the variational problem [6, 10] of the form

$$\min_{u\in V}\ \frac{1}{2}a(u,u)-f(u)+\varphi(\Lambda u), \tag{2.1}$$

where a is a continuous, coercive bilinear form on $X \times X$, $f \in X^*$ is a bounded linear functional on X, and

$$\varphi(v) = I_{v \leq \psi} \qquad \text{(inequality constraint)}, \tag{2.2}$$

$$\varphi(u) = \int_{\Gamma} g(x)|v|\,dx \quad \bigl(\text{weighted } L^1 \text{ norm}\bigr). \tag{2.3}$$

where $v \leq \psi$ is pointwise, i.e., $v(x) \leq \psi(x)$ a.e. $x \in \Gamma$. It will be shown that the necessary optimality of (2.1) is of the form of variational inequalities.

Let Ω be a bounded open set in R^d with boundary $\partial\Omega$ and Λ be the trace operator of $H^1(\Omega)$ onto $\Gamma \subset \partial\Omega$. The elastic contact problem is formulated as (2.1), i.e., $u \in X = H^1(\Omega)^d$ represents the deformation vector and the bilinear form is defined by

$$a(u, \phi) = \int_{\Omega} \sigma : \epsilon(\phi)\,dx,$$

where ϵ is the strain tensor is defined by

$$\epsilon(\phi) = \frac{1}{2}\left(\frac{\partial\phi_i}{\partial x_j} + \frac{\partial\phi_j}{\partial x_i}\right),$$

and σ is the stress tensor. We assume Hooke's law for the strain stress-strain relationship;

$$\sigma = 2\mu\epsilon + \lambda\,\mathrm{tr}\,\epsilon I. \tag{2.4}$$

The functional $f \in X^*$ is defined by $f(u) = \int_{\Gamma_0} f u\,ds$ at an applied body force f at a part Γ_0 of the boundary and let $H = L^2(\Gamma_c)$ where Γ_c is the contact boundary with the solid.

The Signorini problem is of the first kind, i.e., (2.1) subject to $n \cdot u \leq \psi$. The Coulomb friction problem is of the second kind in which $g = \mathcal{F}$ is a Tresca friction coefficient and Λ is the trace operator of $H_1(\Omega)$ to the contact surface Γ_c [8, 15]. The Bingham fluid and solid model [6] for the plastic deformation can be formulated as (2.1) of the second kind with Λu being the strain tensor of the deformation u.

2.2 Inverse Problems and Image Analysis

Inverse problems can be cast as minimizing

$$\phi\bigl(|K(y) - b|\bigr) + \frac{\alpha}{2}|Dy|^2 + \beta\varphi(\Lambda y),$$

where $y \in L^2(\Omega)$ represents the image or the distribution, $K : X \to Y$ is a (nonlinear) convolution operator, and $b \in Y$ is a measurement. The functional ϕ is the fit-to-data criterion [9]. For example, consider the inverse medium problem for the Helmholtz equation

$$\Delta u + n^2 k^2 u = 0, \quad x \in R^d,$$

where k is a given wave number, $u = u^{scat} + u^{inc}$ and n is the refraction index of the medium. Or equivalently, the total field u satisfies

$$u(x) = u^{inc}(x) + \int_\Omega G(x,\omega)\big(n^2 - 1\big)k^2 u(\omega)\, d\omega, \tag{2.5}$$

where Ω is a domain that contains inhomogeneities $\{n^2 > 1\}$ and $G(x,y)$ is the fundamental solution of the homogeneous Helmholtz equation. We assume the plane wave incident, $u^{inc} = e^{ik(x\cdot d)}$ with directions d. The problem is to determine the index n^2 based on measurements u^{scat} at a near field $\Gamma = \{|x| = 5\lambda\}$, where λ is the wave-length. Thus, in this case $y = n^2 - 1$ is the unknown, $K(y) = u^{scat}$ where u solves (2.5) and $b \in Y = L^2(\Gamma)$ is the measurements. For the L_2 regularization term, D is a differential operator, and for the L_1 criterion,

$$\varphi(v) = \int_\Omega |v|\, dx.$$

Λ is the natural injection for the sparsity imaging, while $\Lambda = \nabla$ for the BV-regularization, and $\Lambda = \Delta$ for the nonlinear filter. It is very essential to select proper regularization functionals to obtain the enhanced and robust reconstruction that captures the specific features and properties.

In the case of the registration of the images I_0 and I_1 defined on the domain Ω, find vector field u that minimizes

$$\int_\Omega \Phi\big(\big|I_0\big(x + u(x)\big) - I_1(x)\big|\big)\, dx + \eta\varphi(\Lambda u) \tag{2.6}$$

where $\Phi : R \to R^+$ is a continuous and convex function (for example, Huber function for the robust statistics), Λu is the strain tensor of u, and the selection regularization functional φ plays a significant role and assumes a priori knowledge of the transformation, e.g. the divergence free constraint and elastic deformation energy [1].

2.3 Control and Design Problems

In the case of the optimal control problem [10]

$$\min \int_0^T \ell\big(x(t)\big) + h\big(u(t)\big)\, dt$$

subject the dynamical constraint:

$$\frac{d}{dt}x(t) = f\big(x(t), u(t)\big) \text{ in } X_0^*, \quad u \in \mathcal{C},$$

where the state function $x(t) \in X_0$, a Hilbert space, and control function $u(t) \in \mathcal{C} \subset X$. For example, the parabolic control problem is given by

$$\left\langle \frac{d}{dt}x(t), \phi \right\rangle + a\big(x(t), \phi\big) + \big(u(t), B^*\phi\big)_U, \quad \text{for all } \phi \in X_0,$$

where $a(\cdot,\cdot)$ is a bounded bilinear form on $X_0 \times X_0$ and $B \in \mathcal{L}(X, X_0^*)$ is the input operator. If we eliminate $x(t) \in C(0,T;X_0)$ as a function of $u(t) \in X = L^2(0,T;U)$ by $x = \Phi(u)$ it reduces to (1.1) with

$$F(u) = \int_0^T \ell\big(\Phi(u)\big)(t)\,dt, \quad \text{and} \quad \varphi(\Lambda u) = \int_0^T h\big(u(t)\big)\,dt.$$

Similarly, in the case of the structural optimization for the elliptic equation;

$$\min \quad \phi\big(u(a)\big) + \eta\varphi(\Lambda a) \quad \text{over } a \in \mathcal{C} \tag{2.7}$$

where $u = u(a) \in H_0^1(\Omega)$ satisfies $-\nabla \cdot (a\nabla u) = f$ for example, and ϕ is the material performance and $\varphi(\Lambda a)$ is the complexity measure of design a. For the topological optimization $\mathcal{C}$ is the binary constraint and $\varphi(a) = BV(a)$.

In general we consider the constrained minimization of the form;

$$\min \quad F(x) + H(u) \quad \text{subject to} \quad E(x,u) = 0, \quad (x,u) \in \mathcal{C}.$$

Here, $E(x,u) = 0$ in X_0^* for $x \in X_0$ and $u \in \mathcal{C}$ is the equality constraint and we use the Lagrange formulation;

$$L(x,u,p) = F(x) + H(u) + \big\langle p, E(x,u)\big\rangle_{X_0 \times X_0^*}.$$

Assuming E and F are C^1 and $E_x(x,u) : X \times \mathcal{C} \to X_0^*$ is surjective, we have the necessary optimality;

$$\begin{cases} E_x(x,u)^* p + F'(x) = 0, \\ u = \underset{v \in \mathcal{C}}{\operatorname{argmin}} \big\{H(v) + \big\langle p, E(x,v)\big\rangle\big\}. \end{cases}$$

The second equation is the optimality condition and the functional H is nonsmooth in general. Based on this optimality condition and our nonsmooth methods, we develop iterative algorithms [10].

2.4 Binary and Mixed Integer Program

Consider the binary and mixed integer programming

$$\min \quad F(x) \quad \text{subject to} \quad E(x,u)=0, \quad \text{and} \quad u\in\mathcal{C}=\{0,1,\dots,N\}. \tag{2.8}$$

One can formulate (2.8) as

$$\min \quad F(x)+\beta(Qu,u)+cW(u) \quad \text{subject to} \quad E(x,u)=0$$

where $c\gg 1$ and $\beta\ge 0$ and Q is a generator of Markov process for $u\in\mathcal{C}$ and

$$W(u)=\min\big(|u|,|u-1|,\dots,|u-N|\big).$$

2.5 Stochastic Control Problem

Consider the stochastic control problem;

$$\min \quad E^{0,x}\left[\int_0^\tau e^{-ct}\big(f(X_t)+|u_t|\big)\,dt\right]$$

subject to

$$dX_t=\big(b(X_t)+Gu_t\big)\,dt+\sigma\,dB_t \tag{2.9}$$

over $u_t\in\{\mathcal{F}_t$-adapted integrable control, $|u_t|\le\gamma$, a.e.$\}$ and $\tau>0$ is the exit time of the Ito diffusion X_t from a domain Ω. Let $\mathcal{L}$ is the generator of X_t

$$\mathcal{L}\phi=\frac{1}{2}a_{i,j}(x)\phi_{x_ix_j}+b_i(x)\phi_{x_i}.$$

Let V be the solution to the Hamilton–Jacobi–Bellman equation:

$$\mathcal{L}V-cV+f+\gamma\min\big(0,1-\big|G^t\nabla V\big|\big)=0, \quad u=0 \quad \text{at } x\in\partial\Omega.$$

Then, one has a feedback solution (Markov control);

$$\alpha_t=\alpha(x_t)=\begin{cases}-\gamma\frac{G^t\nabla V}{|G^t\nabla V|}, & x\in\{|G^t\nabla V|\ge 1\},\\ 0, & \text{otherwise.}\end{cases}$$

3 Nonsmooth Optimization and Lagrange Multiplier

In this section we present the Lagrange multiplier theory for the nonsmooth optimization (1.1). We describe the Lagrange multiplier theory to deal with the nonsmoothness of φ. To briefly explain our approach let $u,\lambda\in H$ and $c>0$, and define the family of generalized Yosida–Moreau approximations $\varphi_c(u,\lambda)$ by

$$\varphi_c(u,\lambda) = \inf_{v\in H}\left\{\varphi(u-v) + (\lambda, v)_H + \frac{c}{2}|v|_H^2\right\}. \tag{3.1}$$

Then, the augmented Lagrangian formulation [10, 11, 13] of (1.1) is given by

$$\min_{y\in C} \quad L_c(y,\lambda) = f(x) + \varphi_c(\Lambda y, \lambda). \tag{3.2}$$

Here $\varphi_c(u,\lambda)$ is continuously Fréchet differentiable with respect to $u \in H$ and if $y_c \in \mathcal{C}$ denotes the solution to (3.2), then it satisfies

$$\begin{cases} \langle f'(y_c) + \Lambda^*\lambda_c, y - y_c\rangle_{X^*,X} \geq 0, & \text{for all } y \in \mathcal{C}, \\ \lambda_c = \varphi_c'(\Lambda y_c, \lambda_c). \end{cases}$$

Under appropriate conditions [10, 11, 13] the pair $(y_c, \lambda_c) \in \mathcal{C} \times H$ has a (strong-weak) cluster point $(\bar{y}, \bar{\lambda})$ as $c \to \infty$ such that $\bar{y} \in \mathcal{C}$ is the minimizer of (1.1) and that $\bar{\lambda} \in H$ is a Lagrange multiplier in the sense that

$$\langle f'(\bar{y}) + \Lambda^*\bar{\lambda}, y - \bar{y}\rangle_{X^*,X} \geq 0, \quad \text{for all } y \in \mathcal{C}, \tag{3.3}$$

with the complementarity condition

$$\bar{\lambda} = \varphi_c'(\Lambda\bar{y}, \bar{\lambda}), \quad \text{for each } c > 0. \tag{3.4}$$

System (3.3)–(3.4) is defined for the primal-dual variable $(\bar{y}, \bar{\lambda})$. The advantage here is that the frequently employed differential inclusion $\bar{\lambda} \in \partial\varphi(\Lambda\bar{y})$ is replaced by the equivalent nonlinear equation (3.4). In many applications, the convex conjugate functional φ^* of φ is given by

$$\varphi^*(v) = I_{K^*}(v),$$

where K^* is a closed convex set in H and I_S is the indicator function of a set S. It is shown in [10, 11, 13] that (3.4) is equivalent to

$$\bar{\lambda} = \mathrm{Proj}_{K^*}(\bar{\lambda} + c\Lambda\bar{y}), \tag{3.5}$$

which is the basis of our approach.

3.1 Inequality Constraint Optimization

Let X be a Hilbert space and consider

$$\min \quad J(y) \quad \text{subject to} \quad y \in \mathcal{C} \tag{3.6}$$

where $J : X \to R$ is C^1 and $\mathcal{C}$ is a closed convex set in X. If $y^* \in \mathcal{C}$ minimizes J over $\mathcal{C}$, then we have the necessary optimality

$$\big(J'(y^*), y - y^*\big) \geq 0 \quad \text{for all } y \in \mathcal{C}. \tag{3.7}$$

In this case we let $f = J$ and Λ be the natural injection and $\varphi = I_{\mathcal{C}}$. If $H = L^2(\Omega)$ and $\mathcal{C}$ is the bilateral constraint $\{y \in X : \phi \leq \Lambda y \leq \psi\}$, the complementarity (3.5) implies that for $c > 0$

$$\begin{cases} J'(y^*) + \Lambda^* \mu = 0, \\ \mu = \max\big(0, \mu + c(\Lambda y - \psi)\big) + \min\big(0, \mu + c(\Lambda y - \phi)\big) \quad \text{a.e.} \end{cases} \tag{3.8}$$

If either $\phi = -\infty$ or $\psi = \infty$, it is unilateral constrained and defines the generalized obstacle problem. The cost functional $J(y)$ can represent the performance index for the optimal control and design problem, the fidelity of data-to-fit for the inverse problem, and deformation and restoring energy for the variational problem.

3.2 L^1-Type Optimization

Let $H = L^2(\Omega)$ and $\varphi(v) = \int_\Omega |v|\, dx$, i.e., consider

$$\min \quad f(y) + \int_\Omega |\Lambda y|_2 \, dx. \tag{3.9}$$

In this case

$$K^* = \big\{\lambda : |\lambda|_2 \leq 1 \text{ a.e.}\big\}$$

and the complementarity (3.5) implies that

$$\begin{cases} f'(y^*) + \Lambda^* \lambda = 0, \\ \lambda = \dfrac{\lambda + c\Lambda y}{\max(1, |\lambda + c\Lambda y|_2)} \quad \text{a.e.} \end{cases}$$

4 Semismooth Newton Method

In this section we present the semismooth Newton method for the necessary optimality condition. Consider the nonlinear equation $F(y) = 0$ in a Banach space X. The generalized Newton update is given by

$$y^{k+1} = y^k - V_k^{-1} F\big(y^k\big), \tag{4.1}$$

where V_k is a generalized derivative of F at y^k. In the finite dimensional space or for a locally Lipschitz continuous function F let D_F denote the set of points at which F is differentiable. For $x \in X = R^n$ we define $\partial_B F(x)$ as

$$\partial_B F(x) = \Big\{ J : J = \lim_{x_i \to x,\, x_i \in D_F} \nabla F(x_i) \Big\}, \tag{4.2}$$

where D_F is dense by Rademacher's theorem which states that every locally Lipschitz continuous function in the finite dimensional space is differentiable almost everywhere. Thus, we take $V_k \in \partial_B F(y^k)$.

In infinite dimensional spaces notions of generalized derivatives for functions which are not C^1 cannot rely on Rademacher's theorem. Here, instead, we shall mainly utilize a concept of generalized derivative that is sufficient to guarantee superlinear convergence of Newton's method [10]. This notion of differentiability is called Newton derivative and is defined below. We refer to [3, 7, 10, 20] for further discussion of the notions and topics. Let X, Z be real Banach spaces and let $D \subset X$ be an open set.

Definition 4.1 (1) $F\colon D \subset X \to Z$ is called Newton differentiable at x, if there exists an open neighborhood $N(x) \subset D$ and mappings $G\colon N(x) \to \mathcal{L}(X, Z)$ such that

$$\lim_{|h| \to 0} \frac{|F(x+h) - F(x) - G(x+h)h|_Z}{|h|_X} = 0.$$

The family $\{G(y) : y \in N(x)\}$ is called an N-derivative of F at x.

(2) F is called semismooth at y, if it is Newton differentiable at y and

$$\lim_{t \to 0^+} G(y + th)h \quad \text{exists uniformly in } |h| = 1.$$

Semismoothness was originally introduced in [17] for scalar-valued functions. Convex functions and real-valued C^1 functions are examples for such semismooth functions [18, 19] in the finite dimensional space.

For example, if $F(y)(s) = \psi(y(s))$, point-wise, then $G(y)(s) \in \psi_B(y(s))$ is an N-derivative in $L^p(\Omega) \to L^q(\Omega)$ under appropriate conditions [10]. We use often $\psi(s) = |s|$ and $\max(0, s)$ for our model examples.

Suppose $F(y^*) = 0$, Then, $y^* = y^k + (y^* - y^k)$ and

$$\big|y^{k+1} - y^*\big| = \big|V_k^{-1}\big(F\big(y^*\big) - F\big(y^k\big) - V_k\big(y^* - y^k\big)\big)\big| \le \big|V_k^{-1}\big| o\big(\big|y^k - y^*\big|\big).$$

Thus, the semismooth Newton method is q-superlinear convergent provided that the Jacobian sequence V_k is uniformly invertible as $y^k \to y^*$. That is, if one can select a sequence of quasi-Jacobian V_k that is consistent, i.e., $|V_k - G(y^*)|$ as $y^k \to y^*$ and invertible, (4.1) is still q-superlinear convergent.

4.1 Primal–Dual Active Set Method

Since

$$\partial_B \max(0,s)=\{0,1\},\quad \partial_B \min(0,s)=\{-1,0\},\quad \text{at } s=0,$$

the semismooth Newton method for the bilateral constraint (3.8) is of the form of the primal-dual active set method [10, 12]:

Primal–Dual Active Set Method

1. Initialize $y^0\in X$ and $\lambda^0\in H$. Set $k=0$.
2. Set the active set $\mathcal{A}=\mathcal{A}^+\cup\mathcal{A}^-$ and the inactive set $\mathcal{I}$ by

$$\begin{aligned}\mathcal{A}^+&=\big\{x\in\Omega:\mu^k+c\big(\Lambda y^k-\psi\big)>0\big\},\\ \mathcal{A}^-&=\big\{x\in\Omega:\mu^k+c\big(\Lambda y^k-\phi\big)<0\big\},\\ \mathcal{I}&=\Omega\setminus\mathcal{A}.\end{aligned}$$

3. Solve for (y^{k+1},λ^{k+1})

$$\begin{cases} J''\big(y^k\big)\big(y^{k+1}-y^k\big)+J'\big(y^k\big)+\Lambda^*\mu^{k+1}=0,\\ \Lambda y^{k+1}(x)=\psi(x),\quad x\in\mathcal{A}^+,\quad \Lambda y^{k-1}(x)=\phi(x),\\ x\in\mathcal{A}^-,\quad \mu^{k+1}(x)=0,\quad x\in\mathcal{I}.\end{cases}$$

4. Stop, or set $k=K+1$ and return to the second seep.

In general J'' is a closed operator on $L^2(\Omega)$ and the algorithm is formal one unless J'' is bounded operator in $H=L^2(\Omega)$. It involves solving a linear system of the form

$$\begin{pmatrix} A & \Lambda^*_{\mathcal{A}^+} & \Lambda^*_{\mathcal{A}^-}\\ \Lambda_{\mathcal{A}^+} & 0 & 0\\ \Lambda_{\mathcal{A}^-} & 0 & 0\end{pmatrix}\begin{pmatrix} y\\ \mu_{\mathcal{A}^+}\\ \mu_{\mathcal{A}^-}\end{pmatrix}=\begin{pmatrix} f\\ \psi\\ \phi\end{pmatrix}.$$

In order to gain the stability the following one-parameter family of the regularization can be used [10, 14];

$$\mu=\alpha\big(\max\big(0,\mu+c(\Lambda y-\psi)\big)+\min\big(0,\mu+c(\Lambda y-\phi)\big)\big),\quad \alpha\to 1^+.$$

4.2 L^1-Type Optimization

Next, we discuss the case of $\varphi(v)=\int_\Omega |v|_2\,dx$ (L^1-type optimization). The complementarity is reduced to the form

$$\lambda\max\big(1+\epsilon,|\lambda+cv|\big)=\lambda+cv$$

for $\epsilon \geq 0$. It is often convenient (but not essential) to use very small $\epsilon > 0$ to avoid the singularity for the implementation of algorithm. Let

$$\varphi_\epsilon(s) = \begin{cases} \frac{s^2}{2\epsilon} + \frac{\epsilon}{2}, & \text{if } |s| \leq \epsilon, \\ |s|, & \text{if } |s| \geq \epsilon. \end{cases} \tag{4.3}$$

Then, it corresponds to the regularized one of (3.9):

$$\min \quad J_\epsilon(y) = F(y) + \varphi_\epsilon(\Lambda y).$$

The semismooth Newton method is given by

$$\begin{aligned} & F'(y^+) + \Lambda^* \lambda^+ = 0, \\ & \begin{cases} \lambda^+ = \dfrac{v^+}{\epsilon} & \text{if } |\lambda + v| \leq 1 + \epsilon, \\ |\lambda + cv|\lambda^+ + \left(\lambda \left(\dfrac{\lambda + cv}{|\lambda + cv|}\right)^t\right)(\lambda^+ + cv^+) & \\ \quad = \lambda^+ + cv^+ + |\lambda + cv|\lambda & \text{if } |\lambda + cv| > 1 + \epsilon. \end{cases} \end{aligned} \tag{4.4}$$

There is no guarantee that (4.4) is solvable for λ^+ and is stable. In order to obtain the compact and unconditionally stable formula we use the damped and regularized algorithm with $\beta \leq 1$;

$$\begin{cases} \lambda^+ = \dfrac{v}{\epsilon} & \text{if } |\lambda + cv| \leq 1 + \epsilon, \\ |\lambda + cv|\lambda^+ - \beta\left(\dfrac{\lambda}{\max(1, |\lambda|)}\left(\dfrac{\lambda + cv}{|\lambda + cv|}\right)^t\right)(\lambda^+ + cv^+) & \\ \quad = \lambda^+ + cv^+ + \beta|\lambda + cv|\dfrac{\lambda}{\max(1, |\lambda|)} & \text{if } |\lambda + cv| > 1 + \epsilon. \end{cases} \tag{4.5}$$

Here, the purpose of the regularization $\frac{\lambda}{|\lambda| \wedge 1}$ is to automatically constrain the dual variable λ into the unit ball. The damping factor β is automatically selected to achieve the stability. Let

$$d = |\lambda + cv|, \qquad \eta = d - 1, \qquad a = \frac{\lambda}{|\lambda| \wedge 1}, \qquad b = \frac{\lambda + cv}{|\lambda + cv|}, \qquad F = ab^t.$$

Then, (4.5) is equivalent to

$$\lambda^+ = (\eta I + \beta F)^{-1}\big((I - \beta F)(cv^+) + \beta d a\big),$$

where by Sherman–Morrison formula

$$(\eta I + \beta F)^{-1} = \frac{1}{\eta}\left(I - \frac{\beta}{\eta + \beta a \cdot b} F\right).$$

Then,

$$(\eta I+\beta F)^{-1}\beta da=\frac{\beta d}{\eta+\beta a\cdot b}a.$$

Since $F^2=(a\cdot b)F$,

$$(\eta I+\beta F)^{-1}(I-\beta F)=\frac{1}{\eta}\left(I-\frac{\beta d}{\eta+\beta a\cdot b}F\right).$$

In order to achieve the stability, we let

$$\frac{\beta d}{\eta+\beta a\cdot b}=1,\quad \text{i.e.,}\quad \beta=\frac{d-1}{d-a\cdot b}\leq 1.$$

Consequently, we obtain a compact Newton step

$$\lambda^+=\frac{1}{d-1}(I-F)\big(cv^+\big)+\frac{\lambda}{|\lambda|\wedge 1}, \tag{4.6}$$

which results in.

Primal–Dual Active Set Method (L^1-Optimization)

1. Initialize: $\lambda^0=0$ and solve $F'(y^0)=0$ for y^0. Set $k=0$.
2. Set inactive set $\mathcal{I}_k$ and active set $\mathcal{A}_k$ by

$$\mathcal{I}_k=\big\{\big|\lambda^k+c\Lambda y^k\big|>1+\epsilon\big\},\quad \text{and}\quad \mathcal{A}_k=\big\{\big|\lambda^k+c\Lambda y^k\big|\leq 1+\epsilon\big\}.$$

3. Solve for $(y^{k+1},\lambda^{k+1})\in X\times H$:

$$\begin{cases} F'\big(y^{k+1}\big)+\Lambda^*\lambda^{k+1}=0,\\ \lambda^{k+1}=\dfrac{1}{d^k-1}\big(I-F^k\big)\big(c\Lambda y^{k+1}\big)+\dfrac{\lambda^k}{|\lambda^k|\wedge 1}\quad \text{in } \mathcal{A}_k\quad \text{and}\\ \Lambda y^{k+1}=\epsilon\lambda^{k+1}\quad \text{in } \mathcal{I}_k. \end{cases}$$

4. Convergent or set $k=k+1$ and Return to Step 2.

This algorithm is unconditionally stable and is rapidly convergent for our test examples.

Remark 4.2

(1) Note that

$$\big(\lambda^+,v^+\big)=\frac{|v^+|^2-(a\cdot v^+)(b\cdot v^+)}{d-1}+a\cdot v^+, \tag{4.7}$$

which implies stability of the algorithm. In fact, since

$$\begin{cases} F'(y^{k+1}) - F'(y^k) + \Lambda^*(\lambda^{k+1} - \lambda^k), \\ \lambda^{k+1} - \lambda^k = (I - F^k)(c\Lambda y^{k+1}) + \left(\dfrac{1}{|\lambda^k|} - 1\right)\lambda^k, \end{cases}$$

we have

$$\begin{aligned} &\left(F'(y^{k+1}) - F'(y^k), y^{k+1} - y^k\right) + c\left((I - F^k)\Lambda y^{k+1}, \Lambda y^{k+1}\right) \\ &\quad + \left(\frac{1}{|\lambda^k|} - 1\right)\left(\lambda^k, \Lambda y^{k+1}\right). \end{aligned}$$

Supposing

$$F(y) - F(x) - \left(F'(y), y - x\right) \ge \omega|y - x|_X^2,$$

we have

$$F(y^k) + \sum_{j=1}^{k} \omega|y^j - y^{j-1}|^2 + \left(\frac{1}{|\lambda^{j-1}| \wedge 1} - 1\right)\left(\lambda^{j-1}, v^j\right) \le F(y^0),$$

where $v^j = \Lambda y^j$. Note that $(\lambda^k, v^{k+1}) > 0$ implies that $(\lambda^{k+1}, v^{k+1}) > 0$ and thus is inactive at step $k + 1$, pointwise. This fact is a key step for proving a global convergence of the algorithm.

(2) If $a \cdot b \to 1^+$, then $\beta \to 1$. Suppose (λ, v) is a fixed point of (4.6), then

$$\left(1 - \frac{1}{|\lambda| \wedge 1}\left(1 - \frac{c}{d-1}\frac{\lambda + cv}{|\lambda + cv|} \cdot v\right)\right)\lambda = \frac{1}{d-1}v.$$

Thus, the angle between λ and $\lambda + cv$ is zero and

$$1 - \frac{1}{|\lambda| \wedge 1} = \frac{c}{|\lambda| + c|v| - 1}\left(\frac{|v|}{|\lambda|} - \frac{|v|}{|\lambda| \wedge 1}\right),$$

which implies $|\lambda| = 1$. It follows that

$$\lambda + cv = \frac{\lambda + cv}{\max(|\lambda + cv|, 1 + \epsilon)}.$$

That is, if the algorithm converges, $a \cdot b \to 1$ and $|\lambda| \to 1$ and it is consistent.

(3) Consider the substitution iterate:

$$\begin{cases} F'(y^+) + \Lambda^*\lambda^+ = 0, \\ \lambda^+ = \dfrac{1}{\max(\epsilon, |v|)}v^+, \quad v = \Lambda y. \end{cases} \tag{4.8}$$

Note that

$$\left(\frac{v}{|v|}, v^+ - v\right) = \left(\frac{|v^+|^2 - |v|^2 + |v^+ - v|^2}{2|v|}\right) dx$$
$$= \left(|v^+| - |v| - \frac{(|v^+| - |v|)^2 + |v^+ - v|^2}{2|v|}\right).$$

Thus,

$$J_\epsilon\left(v^+\right) \le J_\epsilon(v),$$

where the equality holds only if v^+. This fact can be used to prove the iterative method (4.8) is globally convergent [10, 13]. It also suggests to use the hybrid method; for $0 < \mu < 1$

$$\lambda^+ = \frac{\mu}{\max(\epsilon, |v|)} v^+ + (1-\mu)\left(\frac{1}{d-1}(I-F)v^+ + \frac{\lambda}{\max(|\lambda|, 1)}\right),$$

in order to gain the global convergence property without loosing the fast convergence of the Newton method.

5 ℓ^0 Sparsity Optimization

Let $A \in \mathcal{L}(\ell^2)$, $b \in \ell^2$. In this section we consider ℓ^0 minimization of the form

$$\frac{1}{2}|Ax - b|_2^2 + \beta N_0(x) \tag{5.1}$$

where

$$N_0(x) = \text{number of nonzero elements of } x = \sum |x_k|^0.$$

Here, $|s|^0 = 1$, $s \neq 0$ and $|0|^0 = 0$ for $s \in R$. It can be shown that N_0 is a complete metric.

Theorem 5.1 *Problem* (5.1) *has a solution* $\bar{x} \in \ell^0$ *and the necessary optimality condition is given by*

$$\begin{cases} \bar{x}_i = 0 & \textit{if } \left|(A_i, f_i)\right| < \sqrt{2\beta}|A_i|, \\ (A_i, A\bar{x} - b) = 0 & \textit{if } \left|(A_i, f_i)\right| > \sqrt{2\beta}|A_i|. \end{cases} \tag{5.2}$$

For the second case of (5.2), $|(A_i, f_i)| > \sqrt{2\beta}|A_i|$ *is equivalent to* $|\bar{x}_i| > \frac{\sqrt{2\beta}}{|A_i|}$.

Proof Let x^n be a minimizing sequence of (5.1). Let $N(s) = 0$ if $s = 0$, and otherwise $N(s) = 1$. Since $c' = \ell_1$, there exists a subsequence of $\{N(x^n)\}$ converging ℓ^1 weakly star to $N(\bar{x})$ and x^n converges to $\bar{x}$ in ℓ^2. Thus, $\lim_{n\to\infty} N_0(x^n) =$

$\langle N(x^n), 1\rangle \to \langle N(\bar{x}), 1\rangle = N_0(\bar{x})$. Thus, we have

$$\frac{1}{2}|A\bar{x} - b|^2 + \beta N_0(\bar{x}) \le \inf_{x\in\ell_0} J(x),$$

and $\bar{x}$ is a minimizer.

Suppose $\bar{x} \in \ell^0$ is a minimizer. Then, $\bar{x}_i \in R$ minimizes

$$G(x_i) = \frac{1}{2}|A_i x_i - f_i|^2 + \beta|x_i|^0, \tag{5.3}$$

where

$$A_i = Ae_i \quad \text{and} \quad f_i = b - A\tilde{x}, \quad \tilde{x} = \begin{cases} 0 & k = i, \\ \bar{x}_k & k \ne i. \end{cases}$$

If $x_i = z > 0$ is a minimizer of (5.3), then

$$|A_i|^2 z - (A_i, f_i) = 0, \qquad G(z) = -\frac{(A_i, f_i)^2}{2|A_i|^2} + \beta + G(0). \tag{5.4}$$

If $G(z) < G(0)$, i.e., $|(A_i, f_i)| > \sqrt{2\beta}|A_i|$ then $\bar{x}_i = z$ is a unique minimizer and $|\bar{x}_i| > \frac{\sqrt{2\beta}}{|A_i|}$. If $|(A_i, f_i)| < \sqrt{2\beta}|A_i|$ then $\bar{x}_i = 0$ is the minimizer of (5.3). If $|(A_i, f_i)| = \sqrt{2\beta}|A_i|$, there are two minimizers $0, z$. □

It follows from the proof of Theorem 5.1 that a minimizer to (5.1) is not necessarily unique. (5.2) is equivalently written as

$$\begin{cases} A^*(Ax - b) + \lambda = 0, \quad \lambda_k x_k = 0 \quad \text{for all } k, \\ \lambda_k = 0, \quad \text{if } k \in \mathcal{I} = \left\{k : |\lambda_k + \Lambda_k x_k|^2 > 2\beta\Lambda_k\right\}, \\ x_j = 0, \quad \text{if } j \in \mathcal{A} = \left\{j : |\lambda_j + \Lambda_j x_j|^2 \le 2\beta\Lambda_j\right\}. \end{cases} \tag{5.5}$$

For the nonlinear case (1.2) we define

$$G_i(x) = F(x) - F(\tilde{x}_i), \quad \text{where } \tilde{x}_k = x_k, k \ne i \text{ and } \tilde{x}_i = 0.$$

In the case of (5.1) we have $G_i(x) = \frac{1}{2}(|A_i x_i - f_i|^2 - |f_i|^2)$.

Corollary 5.2 *Let $\bar{x} \in \ell^0$ is a minimizer of (1.2). The necessary optimality is given by*

$$F'(\bar{x}) + \bar{\lambda} = 0, \quad \bar{\lambda}_i\bar{x}_i = 0 \quad \textit{for all } i,$$

where

$$\bar{x}_i = 0 \quad \textit{if } G_i(\bar{x}) \ge \beta \quad \textit{and} \quad \bar{\lambda}_i = 0 \quad \textit{if } G_i(\bar{x}) < \beta.$$

5.1 $\ell^p (0 < p < 1)$-Sparsity and Globally Convergent Iterative Scheme

In this section we derive a global convergent method based on the regularized formulation. Consider ℓ^p, $0 < p < 1$ minimization of the form

$$\frac{1}{2}|Ax-b|_2^2 + \beta N_p(x) \tag{5.6}$$

where

$$N_p(x) = \sum |y_k|^p.$$

Here, $N_p(x)$ is a complete metric and weakly sequentially lower continuous for $p > 0$ [12], and $N_p(x) \to N_0(x)$ as $p \to 0^+$. Any subsequence of minimizers x_p to (5.6) converges to a minimizer of (5.1) as $p \to 0^+$. In order to overcome the singularity near $s = 0$ for $(|s|^p)' = \frac{ps}{|s|^{2-p}}$, for $\epsilon > 0$ consider the regularized problem:

$$J_\epsilon(x) = \frac{1}{2}|Ax-b|^2 + \Psi_\epsilon\big(|x|^2\big), \tag{5.7}$$

where for $t \geq 0$

$$\Psi_\epsilon(t) = \begin{cases} \frac{p}{2}\frac{t}{\epsilon^{2-p}} + (1-\frac{p}{2})\epsilon^p & t \leq \epsilon^2, \\ t^{\frac{p}{2}} & t \geq \epsilon^2. \end{cases}$$

For $\epsilon > 0$, consider the iterative algorithm for the solution to (5.7):

$$A^*Ax^{k+1} + \frac{\beta p}{\max(\epsilon^{2-p}, |x^k|^{2-p})}x^{k+1} = A^*b. \tag{5.8}$$

Multiplying this by $x^{k+1} - x^k$, we obtain

$$\frac{1}{2}\Big(\big(Ax^{k+1}, x^{k+1}\big) - \big(Ax^k, x^k\big) + \big(A\big(x^{k+1}-x^k\big), x^{k+1}-x^k\big)\Big)$$
$$+ \frac{\beta p}{\max(\epsilon^{2-p}, |x^k|^{2-p})}\frac{1}{2}\big(\big|x^{k+1}\big|^2 - \big|x^k\big|^2 + \big|x^{k+1}-x^k\big|^2\big) + \big(A^*b, x^{k+1}-x^k\big).$$

Then,

$$\frac{1}{\max(\epsilon^{2-p}, |x^k|^{2-p})}\frac{p}{2}\big(\big|x^{k+1}\big|^2 - \big|x^k\big|^2\big) = \Psi_\epsilon'\big(\big|x^k\big|^2\big)\big(\big|x^{k+1}\big|^2 - \big|x^k\big|^2\big).$$

Since $t \to \Psi_\epsilon(t)$ is concave, we have

$$\Psi_\epsilon\big(\big|x^{k+1}\big|^2\big) - \Psi_\epsilon\big(\big|x^k\big|^2\big) - \frac{1}{\max(\epsilon^{2-p}, |x^k|^{2-p})}\frac{p}{2}\big(\big|x^{k+1}\big|^2 - \big|x^k\big|^2\big) \leq 0,$$

and thus

$$J_\epsilon\big(x^{k+1}\big)+\frac{1}{2}\big(A\big(x^{k+1}-x^k\big),x^{k+1}-x^k\big)+\frac{\beta p}{\max(\epsilon^{2-p},|x^k|^{2-p})}\frac{1}{2}\big|x^{k+1}-x^k\big|^2$$
$$\le J_\epsilon\big(x^k\big). \tag{5.9}$$

We have the following convergence result:

Theorem 5.3 *For $\epsilon>0$ let $\{x_k\}$ is generated by (5.8). Then, $J_\epsilon(x^k)$ is monotonically non increasing, and x_k converges to the minimizer of J_ϵ defined by (5.7).*

Proof It follows from (5.9) that $|x^k|_\infty<\infty$ and

$$\sum_{k=0}^\infty\big|x^{k+1}-x^k\big|_2^2<\infty,$$

and thus there exists a subsequence of $\{x^k\}$ and $x^*\in\ell^p$ such that

$$\lim_{k\to\infty}x_k=\lim_{k\to\infty}x^{k+1}=x^*.$$

It follows from (5.8) that

$$A^*Ax^*+\frac{\beta p}{\max(\epsilon^{2-p},|x^*|^{2-p})}x^*=A^*b,$$

i.e., x^* minimizes J_ϵ. □

From Theorem 5.3 the proposed iterative method (5.8) can be used as a globalization step for the semismooth Newton method which will be discussed in the next section.

5.2 *Augmented Lagrangian Formulation and Primal–Dual Active Set Method*

In this section we develop the augmented Lagrangian formulation and the primal-dual active set strategy for the sparsity optimization (5.1). Let P be a nonnegative self-adjoint operator P and $\Lambda_k=|A_k|_2^2+\alpha P_{kk}$. Consider the augmented Lagrangian functional

$$L(x,v,\lambda)=\frac{1}{2}|Ax-b|_2^2+\frac{\alpha}{2}(Px,x)+\beta\sum_k|v_k|^0$$
$$+\sum_k\bigg(\frac{\Lambda_k}{2}|x_k-v_k|^2+(\lambda_k,x_k-v_k)\bigg).$$

If A is nearly singular, we use $\alpha > 0$ and the regularization functional (x, Px) to regularize the problem. Given (x, λ), L is minimized at

$$v = \Phi(x, \lambda) = \begin{cases} \frac{\lambda_k + \Lambda_k x_k}{\Lambda_k} & \text{if } |\lambda_k + \Lambda_k x_k|^2 > 2\Lambda_k \beta, \\ 0 & \text{otherwise.} \end{cases}$$

Given (v, λ), L is minimized at x that satisfies

$$A^*(Ax - b) + \alpha Px + \Lambda(x - v) + \lambda = 0,$$

where Λ is diagonal operator with entries Λ_k. Thus, the augmented Lagrangian method [10] uses the update:

$$\begin{cases} A^*\big(Ax^{n+1} - b\big) + \alpha Px^{n+1} + \Lambda\big(x^{n+1} - v^n\big) + \lambda^n = 0, \\ v^{n+1} = \Phi\big(x^{n+1}, \lambda^n\big), \\ \lambda^{n+1} = \lambda^n + \Lambda\big(x^{n+1} - v^{n+1}\big). \end{cases} \tag{5.10}$$

If it converges, i.e. x^n, $v^n \to x$ and $\lambda^n \to \lambda$, then

$$\begin{cases} A^*(Ax - b) + \alpha Px + \lambda = 0, \\ \lambda_k = 0 \quad \text{if } |x_k|^2 > \dfrac{2\beta}{\Lambda_k}, \\ x_k = 0, \quad \text{if } |\lambda_k|^2 \le 2\beta\Lambda_k. \end{cases}$$

That is, (x, λ) satisfies the necessary optimality condition (5.2).

Motivated by the augmented Lagrangian formulation we obtain a primal-dual active set method as follows.

Primal–Dual Active Set Method (Sparsity Optimization)

1. Initialize: $\lambda^0 = 0$ and x^0 is determined by $A^*(Ax^0 - b) + \alpha Px^0 = 0$. Set $n = 0$.
2. Solve for (x^{n+1}, λ^{n+1});

$$A^*\big(Ax^{n+1} - f\big) + \alpha Px^{n+1} + \lambda^{n+1} = 0, \tag{5.11}$$

where

$$\begin{cases} \lambda_k^{n+1} = 0, & \text{if } k \in \big\{k : \big|\lambda_k^n + \Lambda_k x_k^n\big|^2 > 2\beta\Lambda_k\big\}, \\ x_j^{n+1} = 0, & \text{if } j \in \big\{j : \big|\lambda_j^n + \Lambda_j x_j^n\big|^2 \le 2\beta\Lambda_j\big\}. \end{cases} \tag{5.12}$$

3. Convergent or set $k = k + 1$ and Return to Step 2.

Note that

$$\begin{cases} \lambda_k = 0 & \text{if } k \in \big\{k : |\lambda_k + \Lambda_k x_k|^2 > 2\beta\Lambda_k\big\}, \\ x_j = 0 & \text{if } j \in \big\{j : |\lambda_j + \Lambda_j x_j|^2 \le 2\beta\Lambda_j\big\}, \end{cases}$$

provides a complementarity condition for (5.2). Thus, if the active set method converges, then the converged (x, λ) satisfies the necessary optimality (5.5). It is observed that the active set method converges globally for all examples we tested.

Remark 5.4

(1) Since $(x^{n+1}, \lambda^{n+1}) = 0$,

$$\big((A^*A + \alpha P)x^{n+1}, x^{n+1}\big) \le \big(Ax^{n+1}, f\big),$$

and thus $|Ax^{n+1}|^2 + \alpha(Px^{n+1}, x^{n+1})$ is bounded.

(2) For all test examples the primal-dual method converges globally. The following estimate is a key to establish a global convergence for the method. Note that

$$\begin{aligned} 0 &= \big(A^*\big(Ax^{n+1} - b\big) + Px^{n+1} + \lambda^{n+1}, x^{n+1} - x^n\big) \\ &= \frac{1}{2}\big(\big|Ax^{n+1} - b\big|^2 + \alpha\big(x^{n+1}, Px^{n+1}\big)\big) - \frac{1}{2}\big(\big|Ax^n - b\big|^2 + \alpha\big(x_n, Px^n\big)\big) \\ &\quad + \frac{1}{2}\big(\big|A\big(x^{n+1} - x^n\big)\big|^2 + \alpha\big(x^{n+1} - x^n, P\big(x^{n+1} - x^n\big)\big)\big) - \big(\lambda^{n+1}, x^n\big), \end{aligned}$$

where

$$\lambda_k^n = 0 \quad \text{and} \quad \big|x_k^n\big| \le \sqrt{\frac{2\beta}{\Lambda_k}} \quad \Rightarrow \quad x_k^{n+1} = 0.$$

(3) The global convergence of the Primal–Dual algorithm is analyzed in [12] for the case when $Q = A^*A + \alpha P$ is an M-matrix and $f = A^*b > 0$, and under the following uniqueness assumption.

Let (x, λ) be a solution to (5.5) and define

$$\begin{aligned} \mathcal{A} &= \big\{\big|\sqrt{\Lambda}^{-\frac{1}{2}}(\Lambda x + \lambda)\big| \le \Lambda^{\frac{1}{2}}\sqrt{2\beta}\big\}, \\ \mathcal{I} &= \big\{\big|\sqrt{\Lambda}^{-\frac{1}{2}}(\Lambda x + \lambda)\big| > \Lambda^{\frac{1}{2}}\sqrt{2\beta}\big\}. \end{aligned}$$

Theorem 5.5 (Uniqueness) *We assume the strict complementarity; there exists* $\delta > 0$ *such that*

$$\min_{\mathcal{I}} \big|\sqrt{\Lambda}^{-\frac{1}{2}}(\Lambda x + \lambda)\big| - \max_{\mathcal{A}} \big|\sqrt{\Lambda}^{-\frac{1}{2}}(\Lambda x + \lambda)\big| \ge \delta\sqrt{\beta},$$

and Q is diagonally dominant; there exists $0 \le \rho < 1$ *such that*

$$\big|\sqrt{\Lambda}^{-\frac{1}{2}}(Q - \Lambda)\sqrt{\Lambda}^{-\frac{1}{2}}\big|_\infty \le \rho.$$

If $\delta > \frac{2\rho}{1-\rho}$, *then* (5.5) *has a unique solution.*

Proof Assume there exist two pairs (x,λ) and $(\hat{x},\hat{\lambda})$ satisfying the necessary optimality (5.5). Then we have

$$Q(x-\hat{x})+\lambda-\hat{\lambda}=0$$

and

$$\Lambda x+\lambda-(\Lambda\hat{x}+\hat{\lambda})=(\Lambda-Q)(x-\hat{x}).$$

Thus, if we let $S=\{x_k=\hat{x}_k=0\}^c$,

$$\big|\sqrt{\Lambda}^{-\frac{1}{2}}(\Lambda x+\lambda)\big|-\big|\sqrt{\Lambda}^{-\frac{1}{2}}(\Lambda\hat{x}+\hat{\lambda})\big|\le\rho|x-\hat{x}|_S.$$

Since if $x_j=0$, then $|(\Lambda^{-\frac{1}{2}}\lambda)_j|\le\sqrt{2\beta}$ and if $\hat{x}_j=0$, then $|(\Lambda^{-\frac{1}{2}}\hat{\lambda})_j|\le\sqrt{2\beta}$,

$$\big|\sqrt{\Lambda}(x-\hat{x})\big|_S\le\frac{2}{1-\rho}\sqrt{2\beta}$$

and

$$\big|\sqrt{\Lambda}^{-\frac{1}{2}}(\Lambda x+\lambda)\big|_{\{\lambda=0\}}-\big|\sqrt{\Lambda}^{-\frac{1}{2}}(\Lambda\hat{x}+\hat{\lambda})\big|_{\{\hat{x}=0\}}\le\rho\big|\sqrt{\Lambda}(x-\hat{x})\big|_S.$$

It thus follows that $\frac{2\rho}{1-\rho}<\delta$, and we have a contradiction. □

References

1. B. Berkels, C. Kondermann, C.S. Garbe, M. Rumpf, Reconstructing optical flow fields by motion inpainting, in *Energy Minimization Methods in Computer Vision and Pattern Recognition*, Lecture Notes in Computer Science, vol. 5681 (Springer, Berlin, 2009), pp. 388–400
2. K. Bredies, D.A. Lorenz, Minimization of non-smooth, nonconvex functionals by iterative thresholding, Preprint
3. X. Chen, Z. Nashed, L. Qi, Smoothing methods and semismooth methods for nondifferentiable operator equations. SIAM J. Numer. Anal. **38**, 1200–1216 (2000)
4. E.J. Candès, T. Tao, Decoding by linear programming. IEEE Trans. Inf. Theory **51**, 4203–4215 (2005)
5. D.L. Donoho, Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
6. R. Glowinski, J.L. Lions, R. Tremolieres, *Numerical Analysis of Variational Inequalities* (North-Holland, Amsterdam, 1981)
7. M. Hintermüller, K. Ito, K. Kunisch, The primal-dual active set strategy as a semismooth Newton method. SIAM J. Optim. **13**, 665–688 (2002)
8. S. Hüeber, G. Stadler, B. Wohlmuth, A primal-dual active set algorithm for three-dimensional contact problems with Coulomb friction. SIAM J. Sci. Comput. **30**, 572–596 (2008)
9. K. Ito, B. Jin, T. Takeuchi, A regularization parameter for nonsmooth Tikhonov regularization. SIAM J. Sci. Comput. **33**, 1415–1438 (2011)
10. K. Ito, K. Kunisch, *Lagrange Multiplier Approach to Variational Problems and Applications*. SIAM Advances in Design and Control (2008)
11. K. Ito, K. Kunisch, Augmented Lagrangian methods for nonsmooth, convex optimizations in Hilbert spaces. J. Nonlinear Anal. **41**, 591–616 (2000)

12. K. Ito, K. Kunisch, A variational approach to sparsity optimization based on Lagrange multiplier theory, SFB-Report 2011-014, University of Graz (2011)
13. K. Ito, K. Kunisch, Augmented Lagrangian formulation of nonsmooth, convex optimization in Hilbert spaces, in *Control of Partial Differential Equations and Applications*, ed. by E. Casas. Lecture Notes in Pure and Applied Mathematics, vol. 174 (1995), pp. 107–117
14. K. Ito, K. Kunisch, The primal-dual active set method for nonlinear optimal control problems with bilateral constraints. SIAM J. Control Optim. **43**, 357–376 (2004)
15. K. Kunisch, G. Stadler, Generalized Newton methods for the 2D-Signorini contact problem with friction. ESAIM Math. Model. Numer. Anal. **39**, 827–854 (2005)
16. R.M. Leahy, B.D. Jeffs, On the design of maximally sparse beamforming array. IEEE Trans. Antennas Propag. **39**, 1178–1187 (1991)
17. R. Mifflin, Semismooth and semiconvex functions in constrained optimization. SIAM J. Control Optim. **15**, 959–972 (1977)
18. L. Qi, J. Sun, A nonsmooth version of Newton's method. Math. Program. **58**, 353–367 (1993)
19. L. Qi, Convergence analysis of some algorithms for solving nonsmooth equations. Math. Oper. Res. **18**, 227–244 (1993)
20. M. Ulbrich, Semismooth Newton methods for operator equations in function spaces. SIAM J. Optim. **13**, 805–841 (2002)
21. S.J. Wright, R.D. Nowak, M.A.T. Figueiredo, Sparse reconstruction by separable approximation. IEEE Trans. Signal Process. **57**, 2479–2493 (2009)
22. C.A. Zarzer, On Tikhonov regularization with non-convex sparsity constraints. Inverse Probl. **25**, 025006 (2009)

Parareal in Time Intermediate Targets Methods for Optimal Control Problems

Yvon Maday, Mohamed-Kamel Riahi, and Julien Salomon

Abstract In this paper, we present a method that enables to solve in parallel the Euler–Lagrange system associated with the optimal control of a parabolic equation. Our approach is based on an iterative update of a sequence of intermediate targets that gives rise to independent sub-problems that can be solved in parallel. This method can be coupled with the parareal in time algorithm. Numerical experiments show the efficiency of our method.

Keywords Control · Optimization · PDEs · Parareal in time algorithm · Hight performance computing · Parallel Algorithm

Mathematics Subject Classification (2010) Primary 49J20 · Secondary 68W10

1 Introduction

In the last decade, parallelism across the time [3], based on the decomposition of the time domain into time sub-domains has been exploited to accelerate the simulation of systems governed by time dependent partial differential equations [4]. Among others, the parareal in time algorithm [1, 5] or multi-shooting schemes [2] have

Part of this work was funded by the ANR-06-CIS6-007-01 project PITAC.

Y. Maday (✉) · M.-K. Riahi
UMR 7598, Laboratoire Jacques-Louis Lions, UPMC Univ. Paris 06, F-75005, Paris, France
e-mail: maday@ann.jussieu.fr

M.-K. Riahi
e-mail: riahi@ann.jussieu.fr

Y. Maday
Division of Applied Mathematics, Brown University, Providence, RI, USA

J. Salomon
CEREMADE, Université Paris-Dauphine, Pl. du Mal. de Lattre de Tassigny, F-75016,
Paris, France
e-mail: salomon@ceremade.dauphine.fr

K. Bredies et al. (eds.), *Control and Optimization with PDE Constraints*,
International Series of Numerical Mathematics 164,
DOI 10.1007/978-3-0348-0631-2_5, © Springer Basel 2013

shown excellent results. In the framework of optimal control, this approach has been used to control parabolic systems [7, 8].

In this paper, we introduce a new approach to tackle such problems. The strategy we follow is based on the concept of target trajectory that has been introduced in the case of hyperbolic systems in [6]. Because of the non-reversibility of parabolic equations, a new definition of this trajectory is considered. It enables us to define at both end points of each time sub-domains relevant initial conditions and intermediate targets, so that the initial problem is split up into independent optimization sub-problems.

The paper is organized as follows: the optimal control problem is introduced in Sect. 2 and the parallelization setting is described in Sect. 3. The properties of the cost functionals involved in the control problem are studied in Sect. 4. The general structure of our algorithm is given in Sect. 5 and its convergence is proven in Sect. 6. In Sect. 7, we propose a fully parallelized version of our algorithm. Some numerical tests showing the efficiency of our approach are presented in Sect. 8.

In the sequel, we consider the optimal control problem associated with the heat equation on a compact set Ω and a time interval $[0,T]$, with $T>0$. We denote by $\|\cdot\|_\Omega$ the space norm associated with $L^2(\Omega)$, and by $\|\cdot\|_{\Omega_c}$ the L^2-norm corresponding to a sub-domain $\Omega_c\subset\Omega$. Also, we use the notations $\|\cdot\|_v$ (resp. $\|\cdot\|_{v_n}$) and $\langle\cdot,\cdot\rangle_v$ (resp. $\langle\cdot,\cdot\rangle_{v_n}$) to represent the norm and the scalar product of the Hilbert space $L^2(0,T;\Omega_c)$ (resp. $L^2(I';\Omega_c)$), with I' a sub-interval of $[0,T]$). Given a function y defined on the time interval $[0,T]$, we denote by $y_{|I'}$ the restriction of y to I'.

2 Optimal Control Problem

Given $\alpha>0$, let us consider the optimal control problem defined by:

$$\min_{v\in L^2([0,T];L^2(\Omega_c))} J(v), \tag{2.1}$$

with

$$J(v)=\frac{1}{2}\|y(T)-y_{\text{target}}\|_\Omega^2+\frac{\alpha}{2}\int_0^T \|v(t)\|_{\Omega_c}^2\,dt,$$

where y_{target} is a given state in $L^2(\Omega)$. Given $\nu>0$, the state y evolves from y_0 on $[0,T]$ according to

$$\partial_t y-\nu\Delta y=\mathcal{B}v.$$

In this equation, Δ denotes the Laplace operator, v is the control term, applied on Ω_c and $\mathcal{B}$ is the natural injection from Ω_c into Ω. We assume Dirichlet conditions for y on the boundary of Ω. The corresponding optimality system reads as

$$\begin{cases}\partial_t y-\nu\Delta y=\mathcal{B}v & \text{on } [0,T]\times\Omega,\\ y(0)=y_0,\end{cases} \tag{2.2}$$

$$\begin{cases} \partial_t p + \nu \Delta p = 0 \quad \text{on } [0, T] \times \Omega, \\ p(T) = y(T) - y_{\text{target}}, \end{cases} \tag{2.3}$$

$$\alpha v + \mathcal{B}^* p = 0, \tag{2.4}$$

where $\mathcal{B}^*$ is the adjoint operator of $\mathcal{B}$.

Note that for any $\alpha > 0$, the functional J is continuous, α-convex in $L^2(0, T; \Omega_c)$ and consequently the system (2.1) has a unique solution $v^\star$. We denote by $y^\star$, $p^\star$ the associated state and adjoint state through (2.2)–(2.3).

3 Time Parallelization Setting

In this section, we describe the relevant setting for a time parallelized resolution of the optimality system. Consider $N \geq 2$ and a subdivision of $[0, T]$ of the form:

$$[0, T] = \bigcup_{n=0}^{N-1} I_n,$$

with $I_n = [t_n, t_{n+1}]$, $t_0 = 0 < t_1 < \cdots < t_{N-1} < t_N = T$. For the sake of simplicity, we assume here that the subdivision is uniform, i.e. for $n = 0, \ldots, N-1$ we assume that $t_{n+1} - t_n = T/N$; we denote $\Delta T = T/N$. Given a control v and its corresponding state y and adjoint state p by (2.2)–(2.3), we define the *target trajectory* by:

$$\chi = y - p \quad \text{on } [0, T] \times \Omega. \tag{3.1}$$

The trajectory χ is not governed by any known partial differential equation, but reaches $\chi(T) = y_{\text{target}}$ at time T from the second equation in system (2.3), hence its denomination.

For $n = 0, \ldots, N-1$, let us consider the sub-problems

$$\min_{v_n \in L^2(I_n; L^2(\Omega_c))} J_n(v_n), \tag{3.2}$$

with

$$J_n(v_n) = \frac{1}{2} \left\| y_n(t_{n+1}) - \chi(t_{n+1}) \right\|_{\Omega}^2 + \frac{\alpha}{2} \int_{I_n} \left\| v_n(t) \right\|_{\Omega_c}^2 dt,$$

where the function y_n is defined by

$$\begin{cases} \partial_t y_n - \nu \Delta y_n = \mathcal{B} v_n \quad \text{on } I_n \times \Omega, \\ y_n(t_n) = y(t_n). \end{cases} \tag{3.3}$$

Recall that this optimal control problem is parameterized by v through the local target $\chi(t_{n+1})$, and also through y and p. Given n, $0 \leq n \leq N-1$, we note that this sub-problem has the same structure as the original one, and is also strictly convex in $L^2(t_n, t_{n+1}; \Omega_c)$. We denote by $v_n^\star$ its solution.The optimality system associated with this optimization problem is given by (3.3) and the equations

$$\begin{cases} \partial_t p_n + \nu \Delta p_n = 0 \quad \text{on } I_n \times \Omega, \\ p_n(t_{n+1}) = y(t_{n+1}) - \chi(t_{n+1}), \end{cases} \tag{3.4}$$

$$\alpha v_n + \mathcal{B}^* p_n = 0. \tag{3.5}$$

4 Some Properties of J and J_n

The introduction of the target trajectory in the last section is motivated by the following result.

Lemma 1 *Denote by $\chi^\star$ the target trajectory defined by* (3.1) *with $y = y^\star$ and $p = p^\star$ and by $y_n^\star$, $p_n^\star$, $v_n^\star$ the solutions of* (3.3)–(3.5) *with $y = y^\star$ and $\chi = \chi^\star$. One has*: $v_n^\star = v^\star_{|I_n}$.

Proof Thanks to the uniqueness of the solution to the sub-problem, it is enough to show that $v^\star_{|I_n}$ satisfies the optimality system (3.3)–(3.5).

Let us first note that $y^\star_{|I_n}$ obviously satisfies (3.3) with $v_n = v^\star_{|I_n}$. It directly follows from the definition of $\chi^\star$ (see (3.1)), that:

$$p^\star(t_{n+1}) = y^\star(t_{n+1}) - \chi^\star(t_{n+1}),$$

so that $p^\star_{|I_n}$ satisfies (3.4). Finally, (3.5) is a consequence of (2.4). The result follows. □

Let HJ denote the Hessian operator associated with J; there exists a strong connection between the Hessian operators HJ and HJ_n of J and J_n, as indicated in the next lemma.

Lemma 2 *The Hessian operator HJ_n coincides with the restriction of HJ to controls whose time supports are included in* $[t_{N-1}, T]$.

Proof First note that J is quadratic so that HJ is a constant operator. Given an increase $\delta v \in L^2([0,T]; L^2(\Omega_c))$, we have:

$$\langle HJ(\delta v), \delta v \rangle_v = \|\delta y(T)\|^2_\Omega + \alpha \int_0^T \|\delta v(t)\|^2_{\Omega_c} dt,$$

where δy is the solution of

$$\begin{cases} \partial_t \delta y - \nu \Delta \delta y = \mathcal{B} \delta v \quad \text{on } [0,T] \times \Omega, \\ \delta y(0) = 0. \end{cases}$$

Given $1 \leq n \leq N$, consider now an increase $\delta v_n \in L^2(I_n; L^2(\Omega_c))$. One finds in the same way that:

$$\left\langle HJ_n(\delta v_n), \delta v_n \right\rangle_{v_n} = \left\| \delta y_n(t_{n+1}) \right\|_{\Omega}^2 + \alpha \int_{t_n}^{t_{n+1}} \left\| \delta v_n(t) \right\|_{\Omega_c}^2 dt,$$

where δy_n is the solution of

$$\begin{cases} \partial_t \delta y_n - \nu \Delta \delta y_n = \mathcal{B} \delta v_n & \text{on } [t_n, t_{n+1}] \times \Omega, \\ \delta y_n(t_n) = 0. \end{cases} \tag{4.1}$$

Suppose now that $\delta v = 0$ on $[0, t_{N-1}]$, it is a simple matter to check that $\delta y \equiv 0$ over $[0, t_{N-1}]$. The restriction of δy on the interval $[t_{N-1}, T]$ thus satisfies $\delta y(t_{N-1}) = 0$ and is consequently (up to a time translation) the solution of (4.1). □

We end this section with an estimate on these Hessian operators.

Lemma 3 *Given $\delta v \in L^2([0, T]; L^2(\Omega_c))$, one has*:

$$\alpha \int_0^T \left\| \delta v(t) \right\|_{\Omega_c}^2 dt \leq \left\langle HJ(\delta v), \delta v \right\rangle_v \leq \beta \int_0^T \left\| \delta v(t) \right\|_{\Omega_c}^2 dt, \tag{4.2}$$

where $\beta = \alpha + C/\sqrt{2}$, with C the Poincaré's constant associated with $L^2(\Omega)$.

The proof of this result is straightforward. We note that, because of Lemma 2, the Hessian operator HJ_n also satisfies (4.2).

5 Algorithm

We are now in a position to propose a time parallelized procedure to solve (2.2)–(2.4). In what follows, we describe the principal steps of a parallel algorithm named "SITPOC" (Serial Intermediate Targets for Parallel Optimal Control).

Algorithm 4 (SITPOC) Consider an initial control v^0 and suppose that, at step k one knows v^k. The computation of v^{k+1} is achieved as follows:

I. Compute y^k, p^k and the associated target trajectory χ^k according to (2.2), (2.3) and (3.1) respectively.
II. Solve approximately the N sub-problems (3.2) in parallel. For $n = 0, \ldots, N-1$, denote by $\tilde{v}_n^{k+1}$ the corresponding solutions and by $\tilde{v}^{k+1}$ the concatenation of $(\tilde{v}_n^{k+1})_{n=0,\ldots,N-1}$.
III. Define v^{k+1} by $v^{k+1} = (1 - \theta^k)v^k + \theta^k \tilde{v}^{k+1}$, where θ^k is defined in order to minimize $J((1 - \theta^k)v^k + \theta^k \tilde{v}^{k+1})$.

Note that we do not explain in detail here the optimization step (Step II) and rather present a general structure of our algorithm. Because of the strictly convex setting, some steps of, e.g., a gradient method or a small number of conjugate gradient method step can be used.

6 Convergence

The convergence of Algorithm 4 can be guaranteed under some assumptions. In what follows, we denote by ∇J the gradient of J.

Theorem 6.1 *Suppose that the sequence* $(v^k)_{k\in\mathbb{N}}$ *defined in Algorithm 4 satisfies, for all* $k \geq 0$:

$$J\left(v^k\right) \neq J\left(v^\infty\right), \tag{6.1}$$

$$\left\|\nabla J\left(v^k\right)\right\|_v \leq \eta \left\|v^{k+1} - v^k\right\|_v, \tag{6.2}$$

and

$$J\left(v^k\right) - J\left(v^{k+1}\right) \geq \kappa \left\|v^k - v^{k+1}\right\|_v^2, \tag{6.3}$$

for a given $\eta > 0$, $\kappa > 0$. *Then* $(v^k)_{k\in\mathbb{N}}$ *converges at least geometrically with a rate* $(1 - \frac{\kappa\alpha^2}{\beta\eta^2}) \in [0, 1)$ *to the solution of* (2.2)–(2.4).

Note that in the case (6.1) is not satisfied, there exists $k_0 \in \mathbb{N}$ such that $v^{k_0} = v^\infty$ and the optimum is reached in a finite number of steps.

Proof Define the shifted functional

$$\widetilde{J}(v) = J(v) - J\left(v^\star\right),$$

and note that because of the definition of $v^\star$, one has

$$\widetilde{J}(v) = \frac{1}{2}\left\langle HJ\left(v - v^\star\right), v - v^\star\right\rangle_v \leq \frac{\beta}{2}\left\|v - v^\star\right\|_v^2. \tag{6.4}$$

Since J is quadratic, for any $v \in L^2(\Omega_c)$

$$\nabla J(v) = HJ\left(v - v^\star\right),$$

and consequently

$$\left\langle \nabla J(v), v - v^\star\right\rangle_v = \left\langle HJ\left(v - v^\star\right), v - v^\star\right\rangle_v \geq \alpha\left\|v - v^\star\right\|_v^2,$$

so that

$$\left\|v - v^\star\right\|_v \leq \frac{1}{\alpha}\left\|\nabla J(v)\right\|_v. \tag{6.5}$$

Combining (6.4) and (6.5), one gets

$$\forall v \in L^2(\Omega_c), \quad \sqrt{\widetilde{J}(v)} \leq \gamma \left\| \nabla J(v) \right\|_v, \tag{6.6}$$

with $\gamma = \frac{1}{\alpha}\sqrt{\frac{\beta}{2}}$. Since $\widetilde{J}(v^k) - \widetilde{J}(v^{k+1}) = J(v^k) - J(v^{k+1}) \geq 0$, we have:

$$\begin{aligned} \sqrt{\widetilde{J}\big(v^k\big)} - \sqrt{\widetilde{J}\big(v^{k+1}\big)} &\geq \frac{1}{2\sqrt{\widetilde{J}(v^k)}}\big(J\big(v^k\big) - J\big(v^{k+1}\big)\big) \\ &\geq \frac{\kappa}{2\sqrt{\widetilde{J}(v^k)}} \left\| v^k - v^{k+1} \right\|_v^2 && (6.7) \\ &\geq \frac{\kappa}{2\gamma \|\nabla J(v^k)\|_v} \left\| v^k - v^{k+1} \right\|_v^2 && (6.8) \\ &\geq \frac{\kappa}{2\gamma\eta \|v^k - v^{k+1}\|_v} \left\| v^k - v^{k+1} \right\|_v^2 && (6.9) \\ &\geq c \left\| v^k - v^{k+1} \right\|_v, && (6.10) \end{aligned}$$

where $c = \frac{\kappa}{2\gamma\eta}$. Indeed (6.7) follows from (6.3), (6.8) from (6.6) and (6.9) from (6.2). It follows from the monotonic convergence of $\sqrt{\widetilde{J}(v^k)}$ that the sequence v^k is Cauchy, thus converges.

Let us now study the convergence rate. Define $r^k = \sum_{\ell=k}^{+\infty} \|v^{\ell+1} - v^{\ell}\|_v$. Summing (6.10) between k and $+\infty$, we obtain:

$$\sqrt{\widetilde{J}\big(v^k\big)} \geq c r^k.$$

Using again (6.6) and (6.2), one finds that:

$$\eta\gamma\big(r^k - r^{k+1}\big) \geq c r^k. \tag{6.11}$$

Note that this inequality implies that $1 - \frac{c}{\eta\gamma} \geq 0$. Define $C := \frac{c}{\eta\gamma}$, we have $0 < C \leq 1$. Because of (6.11):

$$(1 - C)^{-k} r^k \geq (1 - C)^{-(k+1)} r^{k+1},$$

and the result follows. □

We now give an example where hypothesis (6.2)–(6.3) are satisfied.

Corollary 6.2 *Assume that Step II of Algorithm 4 is achieved using one step of a local gradient method and that at step k, the algorithm is initialized with $v_n^k := v^k_{|I_n}$, then (6.2)–(6.3) are satisfied hence the algorithm converges to the solution of (2.2)–(2.4).*

Proof Because of the assumptions, the optimization step (Step II) reads:

$$\tilde{v}_n^{k+1} = v_n^k - \rho_n^k \nabla J_n\big(v_n^k\big).$$

Since the functionals J_n are quadratic, one has:

$$\rho_n^k = \frac{\|\nabla J_n(v_n^k)\|_{\mathcal{V}_n}^2}{\langle HJ_n(\nabla J_n(v_n^k)), \nabla J_n(v_n^k)\rangle_{\mathcal{V}_n}}.$$

A first consequence of these equalities is that:

$$\big\langle \nabla J_n\big(v_n^k\big), \tilde{v}_n^{k+1} - v_n^k\big\rangle_{\mathcal{V}_n} = -\rho_n^k \big\|\nabla J_n\big(v_n^k\big)\big\|_{\mathcal{V}_n}^2 \leq 0,$$

moreover Lemmas 2 and 3 imply:

$$\frac{1}{\beta} \leq \rho_n^k \leq \frac{1}{\alpha}. \tag{6.12}$$

One can also obtain similar estimates over θ^k. In this view, note first that since the only iteration which is considered uses as directions of descent $\nabla J_n(v_n^k) = \nabla J(v^k)_{|I_n}$. Then:

$$\begin{aligned}\theta^k &= -\frac{\langle \nabla J(v^k), \tilde{v}^{k+1} - v^k\rangle_{\mathcal{V}}}{\langle HJ(\tilde{v}^{k+1} - v^k), \tilde{v}^{k+1} - v^k\rangle_{\mathcal{V}}}\\ &= -\frac{1}{\langle HJ(\tilde{v}^{k+1} - v^k), \tilde{v}^{k+1} - v^k\rangle_{\mathcal{V}}} \sum_{n=0}^{N-1} \frac{1}{\rho_n^k} \big\|\tilde{v}_n^{k+1} - v_n^k\big\|_{\mathcal{V}_n}^2.\end{aligned} \tag{6.13}$$

Using (6.12), one deduces $-\frac{\beta}{\alpha} \leq \theta^k \leq -\frac{\alpha}{\beta}$. Since $\theta^k \leq 0$ then:

$$\frac{\alpha}{\beta} \leq \big|\theta^k\big| \leq \frac{\beta}{\alpha}.$$

This preliminary results will now be used to prove (6.2). We have:

$$\begin{aligned}\big\|v^{k+1} - v^k\big\|_{\mathcal{V}} &= \big|\theta^k\big| \big\|\tilde{v}^{k+1} - v^k\big\|_{\mathcal{V}}\\ &= \big|\theta^k\big| \sqrt{\sum_{n=0}^{N-1} \big\|\tilde{v}_n^{k+1} - v_n^k\big\|_{\mathcal{V}_n}^2}\\ &= \big|\theta^k\big| \sqrt{\sum_{n=0}^{N-1} \big(\rho_n^k\big)^2 \big\|\nabla J_n\big(v_n^k\big)\big\|_{\mathcal{V}_n}^2}.\end{aligned}$$

Thus we can lower and upper bound the contribution $\|v^{k+1} - v^k\|_{\mathcal{V}}$ as follows:

$$\frac{\alpha}{\beta^2} \big\|\nabla J\big(v^k\big)\big\|_{\mathcal{V}} \leq \big\|v^{k+1} - v^k\big\|_{\mathcal{V}} \leq \frac{\beta}{\alpha^2} \big\|\nabla J\big(v^k\big)\big\|_{\mathcal{V}}. \tag{6.14}$$

The variations in the functional between two iterations of our algorithm reads as:

$$\begin{aligned}
J\big(v^k\big) - J\big(v^{k+1}\big) &= \big\langle \nabla J\big(v^k\big), v^k - v^{k+1}\big\rangle_v - \frac{1}{2}\big\langle HJ\big(v^k - v^{k+1}\big), v^k - v^{k+1}\big\rangle_v \\
&= -\theta^k\big\langle \nabla J\big(v^k\big), \tilde{v}^{k+1} - v^k\big\rangle_v \\
&\quad - \frac{1}{2}\big(\theta^k\big)^2\big\langle HJ\big(\tilde{v}^{k+1} - v^k\big), \tilde{v}^{k+1} - v^k\big\rangle_v \\
&\quad \big(\text{one uses (6.13)}\big) \\
&= \frac{1}{2}\frac{(\langle \nabla J(v^k), \tilde{v}^{k+1} - v^k\rangle_v)^2}{\langle HJ(\tilde{v}^{k+1} - v^k), \tilde{v}^{k+1} - v^k\rangle_v} \\
&= -\frac{\theta^k}{2}\big\langle \nabla J\big(v^k\big), \tilde{v}^{k+1} - v^k\big\rangle_v \\
&= \frac{|\theta^k|}{2}\left\langle \nabla J\big(v^k\big), \sum_{n=0}^{N-1}\rho_n^k \nabla J_n\big(v_n^k\big)\right\rangle_v \\
&= \frac{|\theta^k|}{2}\sum_{n=0}^{N-1}\rho_n^k\big\langle \nabla J\big(v^k\big), \nabla J_n\big(v_n^k\big)\big\rangle_v \\
&= \frac{|\theta^k|}{2}\sum_{n=0}^{N-1}\rho_n^k\big\langle \nabla J_n\big(v_n^k\big), \nabla J_n\big(v_n^k\big)\big\rangle_{v_n} \\
&= \frac{|\theta^k|}{2}\sum_{n=0}^{N-1}\rho_n^k\big\|\nabla J_n\big(v_n^k\big)\big\|_{v_n}^2 \\
&\geq \frac{1}{2}\frac{\alpha}{\beta^2}\big\|\nabla J\big(v^k\big)\big\|_v^2 \quad \big(\text{one uses (6.14)}\big) \\
&\geq \frac{1}{2}\frac{\alpha}{\beta^2}\left(\frac{\alpha^2}{\beta}\right)^2\big\|v^{k+1} - v^k\big\|_v^2 \\
&\geq \frac{1}{2}\frac{\alpha^5}{\beta^4}\big\|v^{k+1} - v^k\big\|_v^2.
\end{aligned}$$

Hence, we get (6.3) with $\kappa = \frac{1}{2}\frac{\alpha^5}{\beta^4}$. □

7 Parareal Acceleration

The method we have introduced with Algorithm 4 requires in Step I two sequential resolutions of the evolution equation (2.2) on the whole interval $[0, T]$, which does not fit with the parallel setting. In this section, we make use of the parareal in time algorithm to parallelize the corresponding computations.

7.1 Setting

Let us first recall the main features of the parareal in time algorithm. We consider the example of (2.2). In order to solve also this evolution equation in parallel we use the parareal in time scheme [4], we first introduce two solvers a coarse solver, denoted as $\mathcal{G}$, and a fine solver denoted as $\mathcal{F}$. The fine solver is the solver that has been used in the previous sections. The coarse can be the same one but it is based on a larger discretization time step that we choose here equal to ΔT. We then introduce intermediate initial conditions denoted as λ_n^k at every times $(t_n)_{n=0,\dots,N-1}$ that are updated iteratively (index k). Suppose that these values (λ_n^k) are known at step k. Denote by $\mathcal{G}_n(\lambda_n)$ and $\mathcal{F}_n(\lambda_n)$ coarse and fine solutions of (3.3) at time t_{n+1} with λ_n as initial value at time t_n. The update is done according to the following iteration:

$$\lambda_{n+1}^{k+1} = \mathcal{G}_n\big(\lambda_n^{k+1}\big) + \mathcal{F}_n\big(\lambda_n^k\big) - \mathcal{G}_n\big(\lambda_n^k\big).$$

The idea we follow consists in merging this procedure with Algorithm 4, i.e., doing some parareal iterations at each iteration of our algorithm.

7.2 Algorithm

We now give details on the resulting procedure. Since the evolution equations depend on the control, we replace the notations $\mathcal{G}_n(\lambda_n)$ and $\mathcal{F}_n(\lambda_n)$ by $\mathcal{G}_n(\lambda_n, v_n)$ and $\mathcal{F}_n(\lambda_n, v_n)$ respectively. As we need backward solvers to compute p, see (2.3), we also introduce $\widetilde{\mathcal{G}}_n(\mu_{n+1})$ and $\widetilde{\mathcal{F}}_n(\mu_{n+1})$ to denote coarse and fine solutions of (3.4) at time t_n with μ_{n+1} as initial value for the backward problem (given at time t_{n+1}). Note that these backward solvers $\widetilde{\mathcal{F}}_n$ (resp: $\widetilde{\mathcal{G}}_n$) do not depend on the control.

We describe in the following the principal steps of an enhanced version of the SITPOC algorithm. We name it "PITPOC" as Parareal Intermediate Targets for Parallel Optimal Control.

Algorithm 5 (PITPOC) Denote by $v_n^k = v^k_{|I_n}$. Consider a control $(v_n^0)_{n=0,\dots,N-1}$, initial values $(\lambda_n^0)_{n=0,\dots,N}$ through forward scheme $\lambda_{n+1}^0 = \mathcal{G}_n(\lambda_n^0, v_n^0)$, final values $(\mu_n^0)_{n=1,\dots,N}$ through backward scheme $\mu_n^0 = \widetilde{\mathcal{G}}_n(\mu_{n+1}^0)$.

Suppose that, at step k one knows v^k, $(\lambda_n^k)_{n=0,\dots,N}$ and $(\mu_n^k)_{n=1,\dots,N}$. The computation of v^{k+1}, $(\lambda_n^{k+1})_{n=0,\dots,N}$ and $(\mu_n^{k+1})_{n=1,\dots,N}$ is achieved as follows:

I. Build the target trajectory $(\chi_n^k)_{n=1,\dots,N}$ according to a definition similar to (3.1):

$$\chi_n^k = \lambda_n^k - \mu_n^k.$$

II. Solve approximately the N sub-problems (3.2) in parallel. For $n = 0, \dots, N-1$, denote by $\tilde{v}_n^{k+1}$ the corresponding solutions.

III. Define $\tilde{v}^{k+1}$ as the concatenation of the sequence $(\tilde{v}_n^{k+1})_{n=0,\dots,N-1}$.

IV. Compute $(\tilde{\lambda}_n^{k+1})_{n=0,\dots,N}$, $(\mu_n^{k+1})_{n=1,\dots,N}$ by:

$$\tilde{\lambda}_{n+1}^{k+1} = \mathcal{G}_n\big(\tilde{\lambda}_n^{k+1}, \tilde{v}_n^{k+1}\big) + \mathcal{F}_n\big(\lambda_n^k, \tilde{v}_n^{k+1}\big) - \mathcal{G}_n\big(\lambda_n^k, v_n^k\big),$$

$$\mu_n^{k+1} = \widetilde{\mathcal{G}}_n\big(\mu_{n+1}^{k+1}\big) + \widetilde{\mathcal{F}}_n\big(\mu_{n+1}^k\big) - \widetilde{\mathcal{G}}_n\big(\mu_{n+1}^k\big).$$

V. Define v^{k+1} and $(\lambda_n^{k+1})_{n=0,\dots,N}$

$$v^{k+1} = \big(1-\theta^k\big)v^k + \theta^k\tilde{v}^{k+1},$$

$$\lambda_n^{k+1} = \big(1-\theta^k\big)\lambda_n^k + \theta^k\tilde{\lambda}_n^{k+1},$$

where θ^k is defined in order to minimize

$$\frac{1}{2}\big\|\big(1-\theta^k\big)\lambda_N^k + \theta^k\tilde{\lambda}_N^{k+1} - y_{\text{target}}\big\|_{\Omega}^2 + \frac{\alpha}{2}\int_0^T \big\|\big(1-\theta^k\big)v^k(t) + \theta^k\tilde{v}^{k+1}(t)\big\|_{\Omega_c}^2 dt.$$

VI. Set $k = k+1$ and return to I.

8 Numerical Results

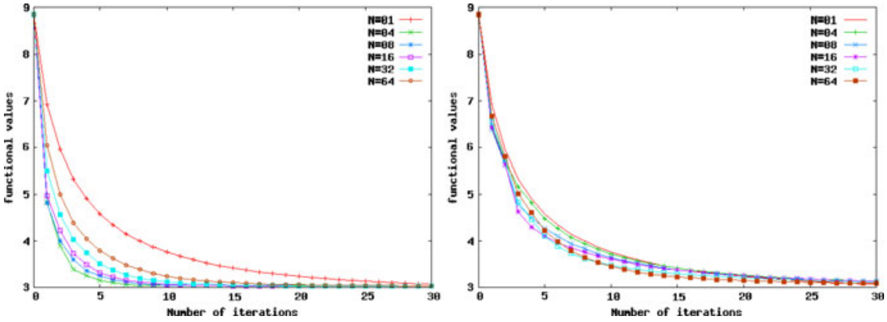
In this section, we test the efficiency of our method and illustrate its robustness. More results can be found in [10].

8.1 *Setting*

We consider a 2D example, where $\Omega = [0,1] \times [0,1]$ and $\Omega_c = [\frac{1}{3}, \frac{2}{3}] \times [\frac{1}{3}, \frac{2}{3}]$. The parameters related to our control problem are $T = 6.4$, $\alpha = 10^{-2}$ and $\nu = 10^{-2}$. The time interval is discretized using a uniform step $\delta t = 10^{-2}$, and an Implicit-Euler solver is used to approximate the solution of (2.2)–(2.3). For the space discretization, we use $\mathbb{P}_1$ finite elements. Our implementation makes use of the freeware `FreeFem` [9] and the parallelization is achieved thanks to the Message Passing Interface (MPI) library. The independent optimization procedures required in Step II are simply carried out using one iterate of an optimal step gradient method.

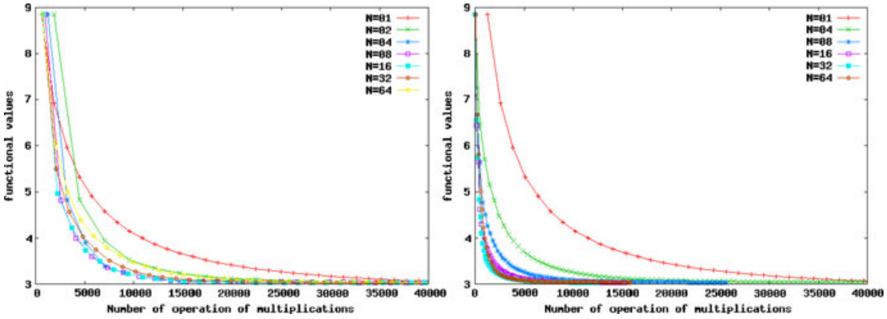
8.2 *Influence of the Number of Sub-intervals*

In this section, Step II of Algorithm 4 and Algorithm 5 are achieved by using one step of an optimal step gradient method. We first test our algorithm by varying



(A) (B)

Fig. 1 Decaying cost functional values according to the iterations count with respect to SITPOC algorithm (**A**) and PITPOC algorithm (**B**)

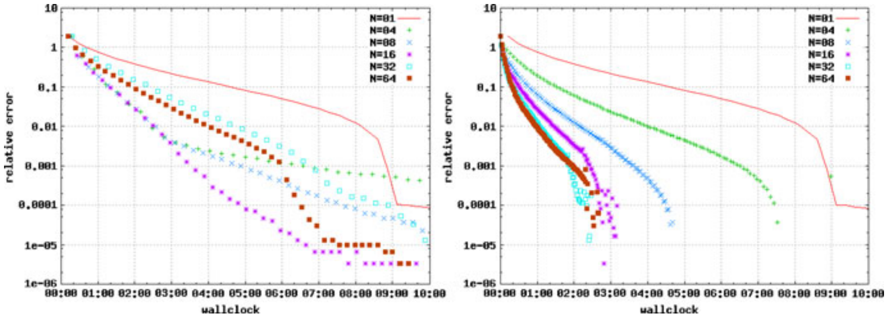


(A) (B)

Fig. 2 Decaying cost functional values according to the multiplication operations count with respect to SITPOC algorithm (**A**) and PITPOC algorithm (**B**)

the number of sub-intervals. The evolution of the cost functional values are plotted with respect to the number of iteration (Fig. 1), the number of matrix multiplication (Fig. 2) and the number of wall-clock time of computation (Fig. 3). We first note that Algorithm 4 actually acts as a preconditioner, since it improves the convergence rate of the optimization process. The introduction of the intermediates targets allows to accelerate the decrease of the functional values, as shown in Fig. 2(A). Note that this property holds mostly for small numbers of sub-intervals, and disappears when dealing with large subdivisions. This feature is lost when considering Algorithm 5, whose convergence does not significantly depend on the number of sub-intervals that is considered, see Fig. 2(B).

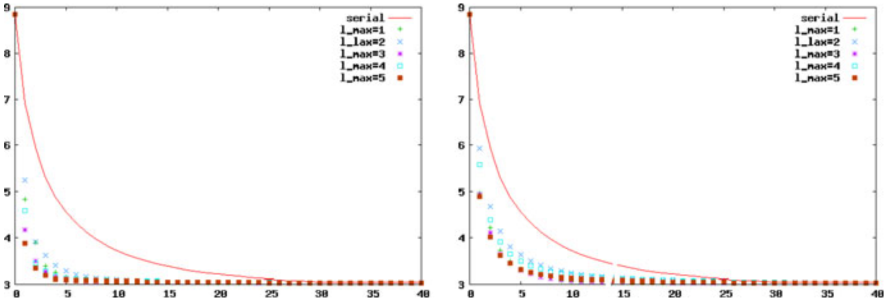
On the contrary, Algorithm 5 achieves a good acceleration when considering the number of multiplications involved in the computations. The corresponding results are shown in Fig. 2, where the parallel operations have been counted only once. We see that Algorithm 5 is close to the full efficiency, since the number of multiplica-



(A)

(B)

Fig. 3 Decaying cost functional values according to elapsed real time with respect to SITPOC algorithm (**A**) and PITPOC algorithm (**B**)



(A)

(B)

Fig. 4 SITPOC algorithm with 4 subdivisions (**A**) and 16 subdivision (**B**): variation of the number of (lower/local) inner-iterations $\ell_{\max}$

tions required to obtain a given value for the cost functional is roughly proportional to $\frac{1}{N}$.

We finally consider the wall-clock time required to carry out our algorithms. As the main part of the operations involved in the computation consists in matrix multiplications, the results we present in Fig. 3 are close to the ones of Fig. 2.

8.3 Influence of the Number of Steps in the Optimization Method in Algorithm 4

We now vary the number $\ell_{\max}$ of steps of the gradient method used in Step II of the Algorithm 4. The results are presented in Fig. 4. Subdivisions of $N = 4$ and $N = 16$ intervals are considered. In both cases, we see that an increase in the number of gradient steps improves the preconditioning feature of that algorithm. However, we also observe that this strategy saturates for large numbers of gradient steps which probably reveals that the sub-problems considered in Step II are practically solved after 5 sub-iterations.

References

1. G. Bal, Y. Maday, A parareal time discretization for non-linear PDEs with application to the pricing of an American put, in *Recent Developments in Domain Decomposition Methods*, Lect. Notes Comput. Sci. Eng. (Springer, Berlin, 2002), pp. 189–202
2. A. Bellen, M. Zennaro, Parallel algorithms for initial value problems for nonlinear vector difference and differential equations. J. Comput. Appl. Math. **25**, 341–350 (1989)
3. K. Burrage, *Parallel and Sequential Methods for Ordinary Differential Equations*. Numerical Mathematics and Scientific Computation (Clarendon Press, Oxford, 1995)
4. J.-L. Lions, Virtual and effective control for distributed systems and decomposition of everything. J. Anal. Math. **80**, 257–297 (2000)
5. J.-L. Lions, Y. Maday, G. Turinici, Résolution d'EDP par un shéma pararréel. C. R. Acad. Sci. Paris, I **332**, 661–668 (2001)
6. Y. Maday, J. Salomon, G. Turinici, Parareal in time control for quantum systems. SIAM J. Numer. Anal. **45**(6), 2468–2482 (2007)
7. Y. Maday, G. Turinici, A parareal in time procedure for the control of partial differential equations. C. R. Math. Acad. Sci. Paris **335**(4), 387–392 (2002)
8. T.P. Mathew, M. Sarkis, C.E. Schaerer, Analysis of block parareal preconditioners for parabolic optimal control problems. SIAM J. Sci. Comput. **32**(3), 1180–1200 (2010)
9. O. Pironneau, F. Hecht, K. Ohtsuka, `FreeFem++-mpi`, http://www.freefem.org
10. M.-K. Riahi, Conception et analyse d'algorithmes parallèles en temps pour l'accélération de simulations numériques d'équations d'évolution. Thèse de doctorat de l'université, Pierre et Marie Curie, Paris 6, July 2012

Hamilton–Jacobi–Bellman Equations on Multi-domains

Zhiping Rao and Hasnaa Zidani

Abstract A system of Hamilton–Jacobi (HJ) equations on a partition of $\mathbb{R}^d$ is considered, and a uniqueness and existence result of viscosity solution is analyzed. While the notion of viscosity solution is by now well known, the question of uniqueness of solution, when the Hamiltonian is discontinuous, remains an important issue. A uniqueness result has been derived for a class of problems, where the behavior of the solution, in the region of discontinuity of the Hamiltonian, is assumed to be irrelevant and can be ignored (see (Camilli, Siconolfi in Adv. Differ. Equ. 8(6):733–768, 2003)). Here, we provide a new uniqueness result for a more general class of Hamilton–Jacobi equations.

Keywords Hamilton–Jacobi equations · Discontinuous Hamiltonian · Viscosity solution · Optimal control

Mathematics Subject Classification (2010) 35F21 · 49L25 · 49L20

1 Introduction

The present work aims at investigating a system of Hamilton–Jacobi–Bellman equations on multi-domains. Consider the repartition of $\mathbb{R}^d$ by disjoint subdomains $(\Omega_i)_{i=1,\dots,m}$ with

$$\mathbb{R}^d = \overline{\Omega}_1 \cup \cdots \cup \overline{\Omega}_m, \qquad \Omega_i \cap \Omega_j = \emptyset \quad \text{for } i \neq j.$$

The work has been partially supported by the European Union under the 7th Framework Programme "FP7-PEOPLE-2010-ITN", GA number 264735-SADCO.

Z. Rao · H. Zidani (✉)
Commands (ENSTA ParisTech, INRIA Saclay), 828, Boulevard des Maréchaux, 91762 Palaiseau Cedex, France
e-mail: Hasnaa.Zidani@ensta-paristech.fr

Z. Rao
e-mail: Zhiping.Rao@ensta-paristech.fr

K. Bredies et al. (eds.), *Control and Optimization with PDE Constraints*,
International Series of Numerical Mathematics 164,
DOI 10.1007/978-3-0348-0631-2_6, © Springer Basel 2013

Consider a collection of Hamilton–Jacobi–Bellman (HJB) equations

$$\begin{cases} -\partial_t u(t,x) + H_i(x, Du(t,x)) = 0, & \text{for } t \in (0,T),\ x \in \Omega_i, \\ u(T,x) = \varphi(x), & \text{for } x \in \Omega_i, \end{cases} \tag{1.1}$$

with the different Hamiltonians H_i satisfying standard assumptions, and where $\phi : \mathbb{R}^d \to \mathbb{R}$ is a Lipschitz continuous function. We address the question to know what condition should be considered on the interfaces (i.e., the intersections of the sets $\overline{\Omega}_i$) in order to get the existence and uniqueness of solution, and also what should be the precise notion of solution.

In order to identify a global solution satisfying (1.1) on each subdomain Ω_i, one can define a global HJB equation with the Hamiltonian H defined on the whole $\mathbb{R}^d$ with $H(x,p) = H_i(x,p)$ whenever $x \in \Omega_i$. However, H can not be expected to be continuous and the definition of H on the interfaces between the subdomains Ω_i is not clear.

The viscosity notion has been introduced by Crandall–Lions to give a precise meaning to the HJ equations with continuous Hamiltonians. This notion has been extended to the discontinuous case by Ishii (see [14]), and later to the case where the Hamiltonian is measurable with respect to the space variable (see [7]). The main difficulty remains the uniqueness of viscosity solution when the Hamiltonian is not continuous.

In [16], a stationary HJ equations with discontinuous Lagrangian have been studied where the Hamiltonian is the type of $H(x,p) + g(x)$ with continuous H and discontinuous g. A uniqueness result is proved under rather restrictive assumptions on g. In [7], the viscosity notion has been extended for the HJ equations with space-measurable Hamiltonians, and a uniqueness result has been established under a transversality assumption. Roughly speaking, this transversality condition amounts saying that the behavior of the solution on the interfaces is not relevant and can be ignored. In the present work, we will consider some more general situations where the transversality condition may not be satisfied. Our aim is to derive some *junction* conditions that have to be considered on the interfaces in order to guarantee the existence and uniqueness of the viscosity solution of (1.1).

Let us mention that the first work dealing with the case where the whole space is separated into two subdomains by one interface has been studied in [4]. In the context, general results on the viscosity sense and uniqueness of solution are analyzed. Even though the problem in [4] considers the problem of a steady equation, the paper shares the same difficulty as the ones we are presenting here. In the present work, our approach is completely different from the one used in [4] and seems to be easy to generalize for two or multi-domains problems. Other papers related to the topic of HJB equations with discontinuous Hamiltonians are [1, 13], where some HJB equations are studied on networks (union of a finite number of half-lines with a single common point), motivated by some traffic flow problems. An inspiring result is the strong comparison principle of [13] leading to the uniqueness result by considering a HJ equation on the junction point.

In the present work, we will investigate the junction conditions on the interfaces. For this, by using the Filippov regularization of the multifunctions F_i, we shall introduce a particular optimal control problem on $\mathbb{R}^d$. The main feature of this control problem is that its value function is solution to the system of (1.1). By investigating the transmission conditions satisfied by the value function on the interfaces between the subdomains Ω_i, we obtain the equations which are defined on the interfaces. Then the system (1.1) is completed by these equations on the interfaces and the existence and uniqueness of solution is guaranteed. No transversality requirement is needed in this paper. The main idea developed here follows the concept of *Essential Hamiltonian* introduced in [5], and provides a new viscosity notion that is quite different from the notion of Ishii [14]. This new definition gives a precise meaning to the transmission conditions between Ω_i and provides the uniqueness of viscosity solution.

The paper is organized as follows. In Sect. 2, the setting of the problem is described and the main results are presented. Section 3 is devoted to the link with optimal control problem and the study of the properties of the value function, and the proofs for the main results are given in Sect. 4.

2 Main Results

2.1 Setting of the Problem

Consider the following structure on $\mathbb{R}^d$: given $m \in \mathbb{N}$, let $\{\Omega_1, \dots, \Omega_m\}$ be a finite collection of C^2 open d-manifolds embedded in $\mathbb{R}^d$. For each $i = 1, \dots, m$, the closure of Ω_i is denoted as $\overline{\Omega}_i$. Assume that this collection of manifolds satisfies the following:

$$\textbf{(H1)} \quad \begin{cases} \text{(i)} \quad \mathbb{R}^d = \bigcup_{i=1}^m \overline{\Omega}_i \quad \text{and} \quad \Omega_i \cap \Omega_j = \emptyset \quad \text{when } i \neq j,\ i, j \in \{1, \dots, m\}; \\ \text{(ii)} \quad \text{Each } \overline{\Omega}_i \text{ is proximally smooth and wedged.} \end{cases}$$

The concepts of proximally smooth and wedged are introduced in [9]. For any set $\Omega \subseteq \mathbb{R}^d$, we recall that $\overline{\Omega}$ is proximally smooth means that the signed distance function to $\overline{\Omega}$ is differentiable on a tube neighborhood of $\overline{\Omega}$. $\overline{\Omega}$ is said to be wedged means that the interior of the tangent cone of $\overline{\Omega}$ at each point of $\overline{\Omega}$ is nonempty. The precise definitions and properties are presented in Appendix B.

Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a given function satisfying:

(H2) φ is a bounded Lipschitz continuous function.

Let $T > 0$ be a given final time, for $i = 1, \dots, m$, consider the following system of Hamilton–Jacobi (HJ) equations:

$$\begin{cases} -\partial_t u(t, x) + H_i(x, Du(t, x)) = 0, & \text{for } t \in (0, T),\ x \in \Omega_i, \\ u(T, x) = \varphi(x), & \text{for } x \in \Omega_i. \end{cases} \tag{2.1}$$

The system above implies that on each d-manifold Ω_i, a classical HJ equation is considered. However, there is no information on the boundaries of the d-manifolds

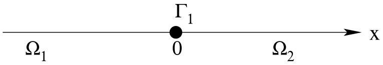


Fig. 1 A multi-domain in 1d

which are the junctions between Ω_i. We then address the question to know what condition should be considered on the boundaries in order to get the existence and uniqueness of solution to all the equations.

In the sequel, we call the singular subdomains contained in the boundaries of the d-manifolds the interfaces. Let $\ell \in \mathbb{N}$ be the number of the interfaces and we denote Γ_j, $j = 1, \dots, \ell$ the interfaces which are also open embedded manifolds with dimensions strictly smaller than d. Assume that the interfaces satisfy the following:

$$\textbf{(H3)} \begin{cases} \text{(i)} & \mathbb{R}^d = (\bigcup_{i=1}^m \Omega_i) \cup (\bigcup_{j=1}^{\ell} \Gamma_j), \quad \Gamma_j \cap \Gamma_k = \emptyset, \quad j \neq k, \ j,k = 1, \dots, \ell; \\ \text{(ii)} & \text{If } \Gamma_j \cap \overline{\Omega}_i \neq \emptyset, \quad \text{then } \Gamma_j \subseteq \overline{\Omega}_i, \quad \text{for } i = 1, \dots, m, \ j = 1, \dots, \ell; \\ \text{(iii)} & \text{If } \Gamma_k \cap \overline{\Gamma}_j \neq \emptyset, \quad \text{then } \Gamma_k \subseteq \overline{\Gamma}_j, \quad \text{for } j,k \in \{1, \dots, \ell\}; \\ \text{(iv)} & \text{Each } \overline{\Gamma}_j \text{ is proximally smooth and relatively wedged.} \end{cases}$$

For any open embedded manifold Γ with dimension $p < d$, $\overline{\Gamma}$ is said to be relatively wedged if the relative interior (in $\mathbb{R}^p$) of the tangent cone of $\overline{\Gamma}$ at each point of $\overline{\Gamma}$ is nonempty, see Appendix B for the precise definition.

Example 1 A simple example is shown in Fig. 1 with $d = 1$, $m = 2$ and $\ell = 1$. Here $\mathbb{R} = \Omega_1 \cup \Gamma_1 \cup \Omega_2$ with

$$\Omega_1 = \{x : x < 0\}, \qquad \Omega_2 = \{x : x > 0\}, \qquad \Gamma_1 = \{0\}.$$

Note that Ω_1, Ω_2 are two one dimensional manifolds, and the only interface is the zero dimensional manifold Γ_1.

Other possible examples in $\mathbb{R}^2$ are depicted in Fig. 2.

We are interested particularly in the HJ equations with the Hamiltonians $H_i : \overline{\Omega}_i \times \mathbb{R}^d \to \mathbb{R}$, $i = 1, \dots, m$ of the following Bellman form: for $(x, q) \in \overline{\Omega}_i \times \mathbb{R}^d$,

$$H_i(x, q) = \sup_{p \in F_i(x)} \{-p \cdot q\},$$

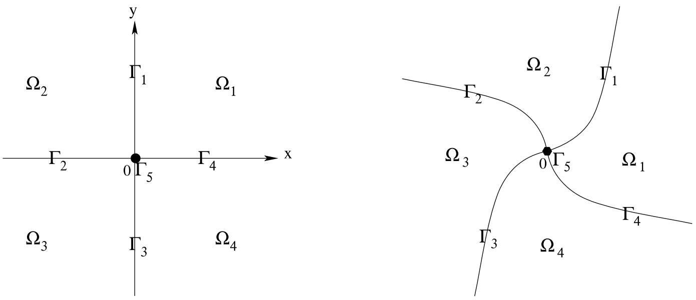


Fig. 2 Other possible examples in $\mathbb{R}^2$

where $F_i : \overline{\Omega}_i \rightsquigarrow \mathbb{R}^d$ are multifunctions defined on $\overline{\Omega}_i$ and satisfy the following assumptions:

$$\textbf{(H4)}\quad \begin{cases} \text{(i)} & \forall x \in \overline{\Omega}_i, \quad F_i(x) \text{ is a nonempty, convex, and compact set;} \\ \text{(ii)} & F_i \text{ is Lipschitz continuous on } \Omega_i \text{ with respect to the} \\ & \text{Hausdorff metric;} \\ \text{(iii)} & \exists \mu > 0 \text{ so that} \quad \max\{|p| : p \in F_i(x)\} \le \mu(1 + \|x\|) \forall x \in \overline{\Omega}_i; \\ \text{(iv)} & \exists \delta > 0 \text{ so that} \quad \forall x \in \overline{\Omega}_i,\ \delta \overline{B(0,1)} \subseteq F_i(x). \end{cases}$$

The hypothesis (H4)(i)–(iii) are classical for the study of HJB equations, whereas (H4)(iv) is a strong controllability assumption. Although this controllability assumption is restrictive, we use it here in order to ensure the continuity of solutions for the system (2.1). The continuity property plays an important role in our analysis, but it can be obtained under weaker assumption than (H4)(iv), see [15].

Remark 2.1 For the simplicity, we define the multifunction F_i on $\overline{\Omega}_i$. In fact, if F_i is only defined on Ω_i and satisfies (H4), it can be extended to the whole $\overline{\Omega}_i$ by its local Lipschitz continuity.

2.2 *Essential Hamiltonian*

The main goal of this work is to identify the junction conditions that ensure the uniqueness of the solution for the HJ system (2.1). In [7], the uniqueness of the solution of space-measurable HJ equations has been studied under some special conditions, called "*transversality*" conditions. Roughly speaking, this transversality condition would mean, in the case of problem (2.1), that the interfaces can be ignored and the behavior of the solution on the interfaces is not relevant. Here we consider the case when no transversality condition is assumed and we analyze the behavior of the solution on the interfaces.

First of all, in order to define a multifunction on the whole $\mathbb{R}^d$, an immediate idea is to consider the approach of Filippov regularization [11] of $(F_i)_{i=1,\dots,m}$. For this consider the multifunction $G : \mathbb{R}^d \rightsquigarrow \mathbb{R}^d$ given by:

$$\forall x \in \mathbb{R}^d, \quad G(x) := \mathrm{co}\Big\{F_i(x) : i \in \{1, \dots, m\}, x \in \overline{\Omega}_i\Big\}.$$

G is the smallest upper semi-continuous (usc) envelope of $(F_i)_{i=1,\dots,m}$ such that $G(x) = F_i(x)$ for $x \in \Omega_i$. Consider the Hamiltonian associated to G:

$$H_G(x, q) = \sup_{p \in G(x)} \{-p \cdot q\}.$$

If $H_G(\cdot, q)$ is Lipschitz continuous, then one could define the HJB equations on the interfaces with the Hamiltonian H_G and the uniqueness result would follow from the classical theory. However, G is not necessarily Lipschitz continuous and the characterization by means of HJB equations is not valid, see [10].

The next step is to define the multifunctions on the interfaces Γ_j. We first recall the notion of tangent cone. For any C^2 smooth $\mathcal{C} \subseteq \mathbb{R}^p$ with $1 \le p \le d$, the tangent cone $\mathcal{T}_{\mathcal{C}}(x)$ at $x \in \mathcal{C}$ is defined as

$$\mathcal{T}_{\mathcal{C}}(x) = \left\{ v \in \mathbb{R}^p : \liminf_{t \to 0^+} \frac{d_{\mathcal{C}}(x + tv)}{t} = 0 \right\},$$

where $d_{\mathcal{C}}(\cdot)$ is the distance function to $\mathcal{C}$. For $j = 1, \ldots, \ell$, we define the multifunction $\widetilde{G}_j : \Gamma_j \rightsquigarrow \mathbb{R}^d$ on the interface Γ_j by

$$\forall x \in \Gamma_j, \quad \widetilde{G}_j(x) := G(x) \cap \mathcal{T}_{\Gamma_j}(x).$$

Note that $\mathcal{T}_{\Gamma_j}(x)$ agrees with the tangent space of Γ_j at x, and the dimension of $\mathcal{T}_{\Gamma_j}(x)$ is strictly smaller than d. On $\widetilde{G}_j$ we have the following regularity result for which the proof is postponed to Appendix A.

Lemma 2.2 *Under the assumptions* (H1), (H2) *and* (H4), $\widetilde{G}_j(\cdot) : \Gamma_j \rightsquigarrow \mathbb{R}^d$ *is locally Lipschitz continuous on* Γ_j.

Through this paper, and for the sake of simplicity of the notations, for $k = 1, \ldots, m + \ell$ we set

$$\mathcal{M}_k = \begin{cases} \Omega_k, & \text{for } k = 1, \ldots, m; \\ \Gamma_{k-m}, & \text{for } k = m + 1, \ldots, m + \ell, \end{cases}$$

and we define a new multifunction $F^{new} : \mathbb{R}^n \rightsquigarrow \mathbb{R}^n$ by

$$F_k^{new}(x) := \begin{cases} F_k(x) & \text{for } x \in \mathcal{M}_k,\ k = 1, \ldots, m; \\ \widetilde{G}_{k-m}, & \text{for } x \in \mathcal{M}_k,\ k = m + 1, \ldots, m + \ell. \end{cases}$$

In all the sequel, we will also need the “essential multifunction” F^E which will be used in the junction conditions:

Definition 2.3 (The essential multifunction) *The essential multifunction* $F^E : \mathbb{R}^d \rightsquigarrow \mathbb{R}^d$ *is defined by*

$$F^E(x) := \bigcup_{k \in \{1, \ldots, m+\ell\}} \left\{ F_k^E(x) : x \in \overline{\mathcal{M}_k} \right\}, \quad \forall x \in \mathbb{R}^d,$$

where $F_k^E : \overline{\mathcal{M}_k} \rightsquigarrow \mathbb{R}^d$ *is defined by*

$$F_k^E(x) = F_k^{new}(x) \cap \mathcal{T}_{\overline{\mathcal{M}_k}}(x), \quad \textit{for } x \in \overline{\mathcal{M}_k}.$$

F^E is called essential velocity multifunction in [5]. According to the definition, $F^E(x)$ is the union of the corresponding inward and tangent directions to each subdomain near x. We note that

$$F^E|_{\mathcal{M}_i} = F_i, \quad \text{for } i = 1, \ldots, m, \quad \text{and} \quad F^E(x) \subseteq G(x), \quad \text{for } x \in \mathbb{R}^d.$$

Example 2 Suppose the following dynamic data for the domain in Example 1:

$$F_1(x) = \left[-\frac{1}{2}, 1\right], \quad \forall x \in \Omega_1, \quad \text{and} \quad F_2(x) = \left[-1, \frac{1}{2}\right], \quad \forall x \in \Omega_2.$$

On this simple example, one can easily see that G and F^E are different on the interface $\{0\}$:

$$G(0) = [-1, 1], \qquad F^E(0) = \left[-\frac{1}{2}, \frac{1}{2}\right].$$

Now, define the "essential" Hamiltonian $H^E : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ by:

$$H^E(x, q) = \sup_{p \in F^E(x)} \{-p \cdot q\}, \quad \forall (x, q) \in \mathbb{R}^d \times \mathbb{R}^d.$$

We point out that on each d-manifold Ω_i, for each $q \in \mathbb{R}^d$

$$H^E(x, q) = H_i(x, q), \quad \text{whenever } x \in \Omega_i.$$

In general, H^E is not Lipschitz continuous with respect to the first variable. Some properties of H^E will be discussed in Sect. 3.

2.3 Main Results

We now state the main existence and uniqueness result.

Theorem 2.4 *Assume that* (H1)–(H4) *hold. The following system*:

$$-\partial_t u(t, x) + H_i\big(x, Du(t, x)\big) = 0, \quad \textit{for } t \in (0, T),\ x \in \Omega_i,\ i = 1, \ldots, m; \tag{2.2a}$$

$$-\partial_t u(t, x) + H^E\big(x, Du(t, x)\big) = 0, \quad \textit{for } t \in (0, T),\ x \in \Gamma_j,\ j = 1, \ldots, \ell; \tag{2.2b}$$

$$u(T, x) = \varphi(x), \quad \textit{for } x \in \mathbb{R}^d, \tag{2.2c}$$

has a unique viscosity solution in the sense of Definition 2.6.

Note that the system (2.2a)–(2.2c) can be rewritten as

$$\begin{cases} -\partial_t u(t, x) + H^E(x, Du(t, x)) = 0, & \text{for } t \in (0, T),\ x \in \mathbb{R}^d, \\ u(T, x) = \varphi(x), & \text{for } x \in \mathbb{R}^d, \end{cases}$$

which is an HJB equation on the whole space with a discontinuous Hamiltonian H^E.

Before giving the definition of viscosity solution, we need the following notion of extended differentials.

Definition 2.5 (Extended differential) *Let $\phi : (0, T) \times \mathbb{R}^d \to \mathbb{R}$ be a continuous function, and let $\mathcal{M} \subseteq \mathbb{R}^d$ be an open C^2 embedded manifold in $\mathbb{R}^d$. Suppose that $\phi \in C^1((0, T) \times \mathcal{M})$. Then we define the differential of ϕ on any $(t, x) \in (0, T) \times \overline{\mathcal{M}}$ by*

$$\nabla_{\overline{\mathcal{M}}}\phi(t, x) := \lim_{x_n \to x, x_n \in \mathcal{M}} \big(\phi_t(t, x_n), D\phi(t, x_n)\big).$$

Note that $\nabla\phi$ is continuous on $(0, T) \times \mathcal{M}$, the differential defined above is nothing but the extension of $\nabla\phi$ to the whole $\overline{\mathcal{M}}$.

Definition 2.6 (Viscosity solution) *Let $u : (0, T] \times \mathbb{R}^d \to \mathbb{R}$ be a bounded local Lipschitz continuous function. For any $x \in \mathbb{R}^d$, let $\mathbb{I}(x) := \{i, x \in \overline{\mathcal{M}_i}\}$ be the index set.*

(i) *We say that u is a supersolution of (2.2a)–(2.2b) if for any $(t_0, x_0) \in (0, T) \times \mathbb{R}^d$, $\phi \in C^1((0, T) \times \mathbb{R}^d)$ such that $u - \phi$ attains a local minimum on (t_0, x_0), we have*

$$-\phi_t(t_0, x_0) + H^E\big(x_0, D\phi(t_0, x_0)\big) \geq 0.$$

(ii) *We say that u is a subsolution of (2.2a)–(2.2b) if for any $(t_0, x_0) \in (0, T) \times \mathbb{R}^d$, any continuous $\phi : (0, T) \times \mathbb{R}^d \to \mathbb{R}$ with $\phi|_{(0,T)\times\mathcal{M}_k}$ being C^1 for any $k \in \mathbb{I}(x)$ such that $u - \phi$ attains a local maximum at (t_0, x_0) on $(0, T) \times \overline{\mathcal{M}_k}$, we have*

$$-q_t + \sup_{p \in F_k^E(x_0)} \{-p \cdot q_x\} \leq 0, \quad \textit{with } (q_t, q_x) = \nabla_{\overline{\mathcal{M}_k}}\phi(t_0, x_0).$$

(iii) *We say that u is a viscosity solution of (2.2a)–(2.2c) if u is both a supersolution and a subsolution, and u satisfies the final condition*

$$u(T, x) = \varphi(x), \quad \forall x \in \mathbb{R}^d.$$

2.4 Comments

The problem (2.1) is formally linked to some hybrid control problems where the dynamics depend on the state region. Theorem 2.4 indicates a new characterization of the value function of hybrid control problems without transition cost. More details are presented in Sect. 3.

Another application related to the addressed problem in this paper is the traffic flow problems where the structure of multi-domains is composed by one-dimensional half-lines and a junction point. On each half-line an HJ equation is imposed to describe the density of the traffic and it is interesting to understand what happens at the junction point. See [13] for more details.

A similar topic with one interface (hyperplane) separating two subdomains has been studied in [4]. The work [4] deals with an infinite horizon problem which leads to the stationary HJB equations with running cost. In this context, a complete analysis of the uniqueness of solutions for (2.1) is provided in [4].

In the present work, we consider a more general situation where the intersection of the domains are interfaces with different dimensions from $d-1$ to zero. In order to focus only on the difficulty arising from this general structure, we consider a time-dependent equations without running cost. The presence of running costs arises to further difficulties that will be addressed in [15].

Optimal control problems on stratified domains have been studied by Bressan–Hong [6] and Barnard–Wolenski [5]. The stratified domains are the multi-domains provided with dynamic data on each subdomain under some structural conditions. The work [5] focuses on the flow invariance on stratified structure. The junction condition established in our work is inspired by the notion of essential dynamics introduced in [5].

3 Link with Optimal Control Problems

Recall that for the classical optimal control problems of the Mayer's type, the value function can be characterized as the unique viscosity solution of the equations of the type (2.1) with Lipschitz continuous Hamiltonians. In our settings of problem, the multifunctions F_i are defined separately on $\overline{\Omega}_i$. A first idea would be to consider the "regularization" of F_i. However, the regularized multifunction G is only usc in general, and this is not enough to guarantee the existence and uniqueness of solution for (2.1). So in our framework, in order to link the Hamilton–Jacobi equation with a Mayer's optimal control problem, we need to well define the global trajectories driven by the dynamics $(F_i)_{i=1,\dots,m}$. Consider the following differential inclusion

$$\begin{cases} \dot{y}(s) \in G(y(s)), & \text{for } s \in (t,T), \\ y(t) = x. \end{cases} \tag{3.1}$$

Since G is usc, (3.1) admits an absolutely continuous solution defined on $[\tau, T]$. For any $(t,x) \in [0,T] \times \mathbb{R}^d$, we denote the set of absolutely continuous trajectories by

$$S_{[t,T]}(x) := \big\{ y_{t,x},\, y_{t,x} \text{ satisfies (3.1)} \big\}.$$

Now consider the following Mayer's problem

$$v(t,x) := \min\big\{ \varphi\big(y(T)\big),\, y(\cdot) \in S_{[t,T]}(x) \big\}. \tag{3.2}$$

Since G is usc and convex, the set $S_{[t,T]}(x)$ of absolutely continuous arcs is compact in $C(t,T;\mathbb{R}^d)$ (see Theorem 1, [2] pp. 60). And then the problem (3.2) has an optimal solution for any $t \in [0,T]$, $x \in \mathbb{R}^d$.

As in the classical case, v satisfies a Dynamical programming principle (DPP).

Proposition 3.1 *Assume that* (H1)–(H3) *hold. Then for any* $(t, x) \in [0, T] \times \mathbb{R}^d$ *the following holds.*

(i) **The super-optimality.** $\exists \bar{y}_{t,x} \in S_{[t,T]}(x)$ *such that*

$$v(t, x) \geq v\big(t + h, \bar{y}_{t,x}(t + h)\big), \quad \textit{for } h \in [0, T - t].$$

(ii) **The sub-optimality.** $\forall y_{t,x} \in S_{[t,T]}(x)$ *such that*

$$v(t, x) \leq v\big(t + h, y_{t,x}(t + h)\big), \quad \textit{for } h \in [0, T - t].$$

An important fact resulting from the assumptions (H2) and (H4)(iv) is the local Lipschitz continuity of the value function v.

Proposition 3.2 *Assume that* (H1)–(H4) *hold. Then the value function* v *is locally Lipschitz continuous on* $[0, T] \times \mathbb{R}^d$.

Proof For any $t \in [0, T]$, we first prove that $v(t, \cdot)$ is locally Lipschitz continuous on $\mathbb{R}^d$. Let $x, z \in \mathbb{R}^d$, without loss of generality, suppose that

$$v(t, x) \geq v(t, z).$$

There exists $\overline{y}_{t,z} \in S_{[t,T]}(z)$ such that

$$v(t, z) = \varphi\big(\overline{y}_{t,z}(T)\big).$$

We set

$$h = \frac{\|x - z\|}{\delta}, \qquad \xi(s) = x + \delta \frac{z - x}{\|z - x\|}(s - t) \quad \text{for } s \in [t, t + h].$$

Note that $\xi(t) = x$, $\xi(t + h) = z$. By the controllability assumption (H4)(iv), we can define the following trajectory

$$\widetilde{y}_{t,x}(s) = \begin{cases} \xi(s), & \text{for } s \in [t, t + h], \\ \overline{y}_{t,z}(s - h), & \text{for } s \in [t + h, T]. \end{cases}$$

By denoting $L_\varphi > 0$ the Lipschitz constant of φ, we have

$$\begin{aligned} v(t, x) - v(t, z) &\leq \varphi\big(\widetilde{y}_{t,x}(T)\big) - \varphi\big(\overline{y}_{t,z}(T)\big) \\ &\leq L_\varphi \big\| \widetilde{y}_{t,x}(T) - \overline{y}_{t,z}(T) \big\| \\ &\leq L_\varphi \big\| \overline{y}_{t,z}(T - h) - \overline{y}_{t,z}(T) \big\| \\ &\leq L_\varphi \|G\| h = \frac{L_\varphi \|G\|}{\delta} \|x - z\|, \end{aligned}$$

where we deduce the local Lipschitz continuity of $v(t, \cdot)$.

Then for $x \in \mathbb{R}^d$, we prove the Lipschitz continuity of $v(\cdot, x)$ on $[0, T]$. For any $t, s \in [0, T]$, without loss of generality suppose that $t < s$. By the super-optimality, there exists $y^{op} \in S_{[t,T]}(x)$ such that

$$v(t, x) = v\big(s, y^{op}(s)\big).$$

Then

$$\big|v(t, x) - v(s, x)\big| = \big|v\big(s, y^{op}(s)\big) - v(s, x)\big| \le L_v \|G\|(s - t),$$

where L_v is the local Lipschitz constant of $v(s, \cdot)$. And the proof is complete. □

Remark 3.3 Assumption (H4)(iv) plays an important role in our proof for the Lipschitz continuity of the value function. However, it is worth mentioning that the Lipschitz continuity can also be satisfied in some cases where (H4)(iv) is not satisfied. In Example 1, if one take $F_1 = F_2$ Lipschitz continuous dynamics, then the value function will be Lipschitz continuous without assuming any controllability property. For multi-domains problems, some weaker assumptions of controllability are analyzed in [15].

The following result analyzes the structure of the dynamics and makes clear the behavior of the trajectories.

Proposition 3.4 *Suppose $y(\cdot) : [t, T] \to \mathbb{R}^d$ is an absolutely continuous arc. Then the following are equivalent.*

(i) *$y(\cdot)$ satisfies* (3.1);
(ii) *For each $k = 1, \dots, m + \ell$, $y(\cdot)$ satisfies $y(t) = x$ and*

$$\dot{y}(s) \in F_k^{new}\big(y(s)\big), \quad \textit{a.e. whenever } y(s) \in \mathcal{M}_k,$$

(iii) *$y(\cdot)$ satisfies*

$$\begin{cases} \dot{y}(s) \in F^E(y(s)) & \textit{for } s \in (t, T), \\ y(t) = x. \end{cases}$$

Proof It is clear that (ii) implies (i) since $F_k^{new}(x) \subseteq G(x)$ whenever $x \in \mathcal{M}_k$. So assume that (i) holds, and let us show that (ii) holds as well.

The proof is essentially the same as in Proposition 2.1 of [5]. For any $k = 1, \dots, m + \ell$, let $J_k := \{s \in [t, T] : y(s) \in \mathcal{M}_k\}$. Without loss of generality, suppose that the Lebesgue measure $mes(J_k) \neq 0$. We set

$$\tilde{J}_k := \big\{s \in J_k : \dot{y}(s) \text{ exists in } G\big(y(s)\big) \text{ and } s \text{ is a Lebesgue point of } J_k\big\}.$$

It is clear that $\tilde{J}_k$ has full measure in J_k. For any $s \in \tilde{J}_k$, then being a Lebesgue point implies that there exists a sequence $\{s_n\}$ such that $s_n \to s$ as $n \to \infty$ with $s \neq s_n \in \tilde{J}_k$ for all n. Since $y(s_n) \in \mathcal{M}_k$, we have

$$\dot{y}(s) = \lim_{n\to\infty} \frac{y(s_n) - y(s)}{s_n - s} \in \mathcal{T}_{\mathcal{M}_k}\big(y(s)\big).$$

Then by the definition of F_k^{new}, we have

$$\dot{y}(s) \in G\big(y(s)\big) \cap \mathcal{T}_{\mathcal{M}_k}\big(y(s)\big) = F_k^{new}\big(y(s)\big), \quad \forall s \in \tilde{J}_k,$$

which proves (ii).

It is clear that (ii) ⇒ (iii) ⇒ (i) since $F_k^{new}(\cdot) \subseteq F^E(\cdot) \subseteq G(\cdot)$, which ends the proof. □

Proposition 3.4 will be very useful in the characterization of the super-optimality and the sub-optimality by HJ equations involving the essential Hamiltonian H^E.

3.1 The Supersolution Property

The following proposition shows the characterization of the super-optimality by the supersolutions of HJ equations. This is a classical result since G is usc.

Proposition 3.5 *Suppose $u : [0,T] \times \mathbb{R}^d \to \mathbb{R}$ is continuous. Then u satisfies the super-optimality if and only if for any $(t_0, x_0) \in (0,T) \times \mathbb{R}^d$, $\phi \in C^1((0,T) \times \mathbb{R}^d)$ such that $u - \phi$ attains a local minimum on (t_0, x_0), we have*

$$-\phi_t(t_0, x_0) + H_G\big(x_0, D\phi(t_0, x_0)\big) \geq 0. \tag{3.3}$$

Proof This is a straightforward consequence of Theorem 3.2 and Lemma 4.3 in [12] (see also [3]). □

Due to the structure of the dynamics G illustrated in Proposition 3.4, it is possible to replace G by F^E to get a more precise HJB inequality since the set of trajectories driven by G or F^E is the same. But the difficulty here is that in general F^E is not usc.

At first, we have the following result concerning the dynamics of the optimal trajectories.

Lemma 3.6 *Let $y(\cdot) \in S_{[t,T]}(x)$ be an absolutely continuous arc along which the value function v satisfies the super-optimality. For any $p \in \mathbb{R}^d$ such that there exists $t_n \to 0^+$ with $\frac{y(t_n)-x}{t_n} \to p$, by denoting $co\ F^E(x)$ the convex hull of $F^E(x)$ we have*

$$p \in co\ F^E(x).$$

The proof of Lemma 3.6 is presented in Appendix A. In the next theorem, we will use the statement of Lemma 3.6 to show that the functions satisfying the super-optimality condition is also a solution to a more precise HJB equation with H^E than the HJB equation (3.3) with the Hamiltonian H_G even if F^E is not usc.

Theorem 3.7 *Suppose $u : [0, T] \times \mathbb{R}^d \to \mathbb{R}$ is continuous and $u(T, x) = \varphi(x)$ for all $x \in \mathbb{R}^d$. u satisfies the super-optimality if and only if u is a supersolution of* (2.2a)–(2.2c), *i.e. for any $(t_0, x_0) \in (0, T) \times \mathbb{R}^d$, $\phi \in C^1((0, T) \times \mathbb{R}^d)$ such that $u - \phi$ attains a local minimum on (t_0, x_0), we have*

$$-\phi_t(t_0, x_0) + \sup_{p \in F^E(x_0)} \left\{-p \cdot D\phi(t_0, x_0)\right\} \geq 0.$$

Proof (⇒) Let $\bar{y}_{t_0,x_0}$ be the optimal trajectory along which u satisfies the super-optimality. Then for any $(t_0, x_0) \in (0, T) \times \mathbb{R}^d$, $\phi \in C^1((0, T) \times \mathbb{R}^d)$ such that $u - \phi$ attains a local minimum on (t_0, x_0), by the same argument in Proposition 3.5, we obtain

$$\frac{1}{h}\big(\phi(t_0, x_0) - \phi\big(t_0 + h, \bar{y}_{t_0,x_0}(t_0 + h)\big)\big) \geq 0,$$

i.e.,

$$\frac{1}{h}\int_0^h \left[-\phi_t\big(t_0 + s, \bar{y}_{t_0,x_0}(t_0 + s)\big) - D\phi\big(t_0 + s, \bar{y}_{t_0,x_0}(t_0 + s)\big) \cdot \dot{\bar{y}}_{t_0,x_0}(t_0 + s)\right] ds \geq 0.$$

Up to a subsequence, let $h_n \to 0^+$ so that $x_n := \bar{y}_{t_0,x_0}(t_0 + h_n)$ satisfies $\frac{x_n - x}{h_n} \to p$ for some $p \in \mathbb{R}^d$. We then get

$$-\phi_t(t_0, x_0) - p \cdot D\phi(t_0, x_0) \geq 0.$$

Lemma 3.6 leads to

$$p \in \text{co } F^E(x_0). \tag{3.4}$$

Then we deduce that

$$-\phi_t(t_0, x_0) + \sup_{p \in \text{co } F^E(x_0)} \left\{-p \cdot D\phi(t_0, x_0)\right\} \geq 0.$$

By the separation theorem

$$-\phi_t(t_0, x_0) + \sup_{p \in F^E(x_0)} \left\{-p \cdot D\phi(t_0, x_0)\right\} \geq 0.$$

(⇐) For any $(t_0, x_0) \in (0, T) \times \mathbb{R}^d$, $\phi \in C^1((0, T) \times \mathbb{R}^d)$ such that $u - \phi$ attains a local minimum on (t_0, x_0), since u is a supersolution, we have

$$-\phi_t(t_0, x_0) + \sup_{p \in F^E(x_0)} \left\{-p \cdot D\phi(t_0, x_0)\right\} \geq 0.$$

Note that $F^E(x_0) \subseteq G(x_0)$, then we deduce that

$$-\phi_t(t_0, x_0) + \sup_{p \in G(x_0)} \left\{-p \cdot D\phi(t_0, x_0)\right\} \geq 0.$$

Then we deduce the desired result by Proposition 3.5. □

3.2 The Subsolution Property

As mentioned before, if G is Lipschitz continuous, one can characterize the sub-optimality by the opposite HJB inequalities:

$$-u_t(t_0, x_0) + H_G\big(x_0, Du(t_0, x_0)\big) \leq 0$$

in the viscosity sense. However, G is only usc on the interfaces. And the characterization using H_G fails because there are dynamics in G which are not "*essential*", which means for some $p \in G(x)$, there does not exist any trajectory coming from x using the dynamic p. For instance in Example 2, at the point 0, $G(0) = [-1, 1]$. Consider the dynamic $p = 1 \in G(0)$, if there exists a trajectory y starting from 0 using the dynamic 1, y goes immediately into Ω_2 and y is not admissible since 1 is not contained in the dynamics F_2.

In the sequel, we consider the essential dynamic multifunction F^E to replace G by eliminating the useless nonessential dynamics. Note that F^E in general is not Lipschitz either. The significant role of F^E is shown in the following result.

Lemma 3.8 *For any $p \in F^E(x)$, there exists $\tau > t$ and a solution $y(\cdot)$ of* (3.1) *which is C^1 on $[t, \tau]$ with $\dot{y}(t) = p$.*

Proof This is a partial result of in [5, Proposition 5.1]. For the convenience of reader, a sketch of the proof is given in Appendix B. □

More precisely, Lemma 3.8 can be rewritten as:

Lemma 3.9 *Let $k \in \{1, \ldots, m + \ell\}$, $x \in \overline{\mathcal{M}_k}$. Then for any $p \in F_k^E(x)$, there exist $\tau > t$ and a trajectory of* (3.1) *$y(\cdot)$ which is C^1 on $[t, \tau]$ with $\dot{y}(t) = p$ and $y(s) \in \overline{\mathcal{M}_k}$ for $s \in [t, \tau]$.*

The following two results give the characterization of sub-optimality by HJB inequalities.

Proposition 3.10 *Let $u : [0, T] \times \mathbb{R}^d \to \mathbb{R}$ be locally Lipschitz continuous and $u(T, x) = \varphi(x)$ for all $x \in \mathbb{R}^d$. Suppose that u satisfies the sub-optimality, then u is a subsolution of* (2.2a)–(2.2c) *in the sense of Definition* 2.6.

Proof Given $(t_0, x_0) \in [0, T] \times \mathbb{R}^d$, for any $k \in \mathbb{I}(x_0)$, $p \in F_k^E(x_0)$, by Lemma 3.9, there exists $h > 0$ and a solution $y(\cdot)$ of (3.1) C^1 on $[t_0, t_0 + h]$ with $\dot{y}(t_0) = p$, $y(t_0) = x_0$ and $y(s) \in \overline{\mathcal{M}_k}$, $\forall s \in [t_0, t_0 + h]$. By the sub-optimality of u

$$u(t_0, x_0) \leq u\big(t_0 + h, y(t_0 + h)\big).$$

For any $\phi \in C^0((0, T) \times \mathbb{R}^d) \cap C^1((0, T) \times \mathcal{M}_k)$ such that $u - \phi$ attains a local maximum at (t_0, x_0) on $(0, T) \times \overline{\mathcal{M}_k}$, we have

$$u\big(t_0 + h, y(t_0 + h)\big) - \phi\big(t_0 + h, y(t_0 + h)\big) \leq u(t_0, x_0) - \phi(t_0, x_0).$$

Then we deduce that

$$\frac{1}{h}\big(\phi(t_0,x_0)-\phi\big(t_0,y(t_0+h)\big)\big)\le 0.$$

By taking $h\to 0$ we have

$$-q_t-p\cdot q_x\le 0,\quad \text{where } p\in F_k^E(x_0),\ (q_t,q_x)\in\nabla_{\overline{\mathcal{M}}_k}\phi(t_0,x_0),$$

i.e.

$$-q_t+\sup_{p\in F_k^E(x_0)}\{-p\cdot q_x\}\le 0. \qquad \square$$

We present a precise example to illustrate that H^E is the proper Hamiltonian for the subsolution characterization of the value function.

Example 3 Consider again the same 1d structure as in Example 1 and Example 2, i.e. $\mathbb{R}=\Omega_1\cup\Omega_2\cup\Gamma_1$ with

$$\Omega_1=(-\infty,0),\qquad \Omega_2=(0,+\infty),\qquad \Gamma_1=\{0\},$$

and the dynamics

$$F_1(x)=\left[-\frac{1}{2},1\right],\quad \forall x\in\Omega_1,\quad \text{and}\quad F_2(x)=\left[-1,\frac{1}{2}\right],\quad \forall x\in\Omega_2.$$

At the point 0, the convexified dynamics $G(0)=[-1,1]$ and the essential dynamics $F^E(0)=[-\frac{1}{2},\frac{1}{2}]$. Let $T>0$ be a given final time and the final cost function $\varphi_2(x)=x$. Then from any initial data $(t,x)\in[0,T]\times\mathbb{R}$, the optimal strategy is to go on the left as far as possible. Thus the value function is given by

$$v_2(t,x):=\min\big\{\varphi_2\big(y_{t,x}(T)\big)\big\}=\begin{cases} x-\frac{1}{2}(T-t) & x\le 0,\\ -\frac{1}{2}(T-t-x) & 0\le x\le T,\\ x-(T-t) & x\ge T-t.\end{cases}$$

At the point $(t,x)=(0,0)$, $\partial_t v_2(0,0)=\frac{1}{2}$, $Dv_2(0,0^-)=1$, $Dv_2(0,0^+)=\frac{1}{2}$, $D^+v_2(0,0)=[\frac{1}{2},1]$. Then we have

$$-\partial_t v_2(0,0)+\max_{p\in F^E(0)}\big\{-p\cdot D^+v_2(0,0)\big\}=0\le 0,$$

while

$$-\partial_t v_2(0,0)+\max_{p\in G(0)}\big\{-p\cdot D^+v_2(0,0)\big\}=\frac{1}{2}>0.$$

We see that the subsolution property fails if we replace F^E by G which is larger.

Proposition 3.7 indicates that any function satisfying the sub-optimality is a subsolution of (2.2a)–(2.2c). The inverse result needs more elaborated arguments. The difficulty arises mainly from handling the trajectories oscillating near the interfaces, i.e. the trajectories cross the interfaces infinitely in finite time which exhibit a type of "Zeno" effect. The proofs of Theorem 3.12 and of Proposition 3.11 contain details on how to construct the "nice" approximate trajectories to deal with Zeno-type trajectories.

At first, we give the following result containing the key fact of Zeno-type trajectories.

Proposition 3.11 *Let u be a Lipschitz continuous subsolution of* (2.2a)–(2.2c). *Suppose $\mathcal{M}_k$ is a subdomain and $\mathcal{M}$ is a union of subdomains with $\mathcal{M}_k \subseteq \overline{\mathcal{M}}$. Assume $\mathcal{M}$ has the following property*: *for every trajectory $y(\cdot)$ of* (3.1) *defined on $[a,b] \subseteq [t,t+h]$ with $y(\cdot) \subseteq \mathcal{M}$, we have*

$$u\big(a, y(a)\big) \le u\big(b, y(b)\big). \tag{3.5}$$

Then for any trajectory $y(\cdot)$ of (3.1) *defined on $[a,b] \subseteq [t,t+h]$ lying totally within $\mathcal{M}_k \cup \mathcal{M}$, we have*

$$u\big(a, y(a)\big) \le u\big(b, y(b)\big).$$

Proof Here we adapt an idea introduced in [5] in a context of stratified control problems. Let $y(\cdot)$ be a trajectory of (3.1) with $y(\cdot) \subseteq \mathcal{M}_k \cup \mathcal{M}$ satisfying (3.5). Without loss of generality, suppose that $y(a) \in \mathcal{M}_k$ and $y(b) \in \mathcal{M}_k$. By (H3), we have $\mathcal{M}_k \cap \mathcal{M} = \emptyset$. Let $J := \{s \in [a,b] : y(s) \notin \mathcal{M}_k\}$, which is an open set and so can be written as

$$J = \bigcup_{n=1}^{\infty} (a_n, b_n)$$

where the intervals are pairwise disjoint. For a fixed p, we set

$$J_p := \bigcup_{n=1}^{p} (a_n, b_n),$$

which after re-indexing can be assumed to satisfy

$$b_0 := a \le a_1 < b_1 \le a_2 < b_2 \le \cdots \le a_p < b_p \le a_{p+1} := b.$$

Choose p sufficiently large so that

$$meas(J \backslash J_p) < \frac{r}{2e^{LT}\|G\|},$$

where $\|G\|$ is an upper bound of the norm of any velocity that may appear, and $r > 0$ is given by

$$r := \inf_{\substack{s\in[b_0,b] \\ w\in\overline{\mathcal{M}}_k\setminus\mathcal{M}_k}} \big\|y(s) - w\big\|.$$

For $n = 1, \ldots, p$, $y(s) \in \mathcal{M}$ for $s \in (a_n, b_n)$. Let $\varepsilon > 0$ small enough such that $[a_n + \varepsilon, b_n - \varepsilon] \subseteq (a_n, b_n)$, then by (3.5)

$$u\big(a_n + \varepsilon, y(a_n + \varepsilon)\big) \leq u\big(b_n - \varepsilon, y(b_n - \varepsilon)\big).$$

Taking $\varepsilon \to 0$ and by the continuity of u and $y(\cdot)$, we deduce that

$$u\big(a_n, y(a_n)\big) \leq u\big(b_n, y(b_n)\big).$$

Next we need to deal with $y(\cdot)$ restricted to $[b_n, a_{n+1}]$. For $n = 0, \ldots, p$, by Proposition 3.4 $\dot{y}(s) \in F_k^{new}(y(s))$ for almost all $s \in [b_n, a_{n+1}]\setminus J$. For $n = 0, \ldots, p$, set $\varepsilon_n := meas([b_n, a_{n+1}] \cap J)$, and note that $\sum_{n=0}^{p} \varepsilon_n = meas(J\setminus J_p)$. We calculate how far $y(\cdot)$ is from a trajectory lying in $\mathcal{M}_k$ with dynamics F_k^{new} by

$$\xi_n := \int_{b_n}^{a_{n+1}} \operatorname{dist}\big(\dot{y}(s), F_k^{new}\big(y(s)\big)\big)ds \leq 2\|G\|\varepsilon_n.$$

By the Filippov approximation theorem (see [8, Theorem 3.1.6] and also [9, Proposition 3.2]), there exists a trajectory $z_n(\cdot)$ of F_k^{new} defined on the interval $[b_n, a_{n+1}]$ that lies in $\mathcal{M}_k$ with $z_n(b_n) = y(b_n)$ and satisfies

$$\big\|z_n(a_{n+1}) - y(a_{n+1})\big\| \leq e^{L(a_{n+1}-b_n)}\xi_n \leq 2\|G\|e^{L(a_{n+1}-b_n)}\varepsilon_n. \tag{3.6}$$

Since u is subsolution of (2.2a)–(2.2c), then for any $x \in \mathcal{M}_k$, note that $F_k^{new}(x) \subseteq \mathcal{T}_{\mathcal{M}_k}(x)$ and $\mathcal{T}_{\mathcal{M}_k}(x) = \mathcal{T}_{\overline{\mathcal{M}}_k}(x)$ by Definition 2.6

$$-\partial_t\phi(t,x) + \sup_{p\in F_k^{new}(x)} \big\{-p\cdot D\phi(t,x)\big\} \leq 0 \tag{3.7}$$

with $\phi \in C^0((0,T) \times \mathbb{R}^d) \cap C^1((0,T) \times \mathcal{M}_k)$ and $u - \phi$ attains a local maximum at (t,x) on $(0,T) \times \mathcal{M}_k$. Since $z_n(\cdot)$ lies in $\mathcal{M}_k$ on $[b_n, a_{n+1}]$ driven by the Lipschitz dynamics F_k^{new}, then (3.7) implies that the sub-optimality of u is satisfied on $z_n(\cdot)|_{[b_n,a_{n+1}]}$, i.e.

$$u\big(b_n, z_n(b_n)\big) \leq u\big(a_{n+1}, z_n(a_{n+1})\big).$$

Then by (3.6) we have

$$\begin{aligned} u\big(b_n, y(b_n)\big) = u\big(b_n, z_n(b_n)\big) &\leq u\big(a_{n+1}, z_n(a_{n+1})\big) \\ &\leq u\big(a_{n+1}, y(a_{n+1})\big) + 2L_u\|G\|e^{L(a_{n+1}-b_n)}\varepsilon_n. \end{aligned}$$

We set $\varepsilon^p := meas(J\backslash J_p)$, and we deduce that

$$\begin{aligned} u\big(a, y(a)\big) &\le u\big(a_1, y(a_1)\big) + 2L_u\|G\|e^{L(a_1-b_0)}\varepsilon_1 \\ &\le u\big(a_2, y(a_2)\big) + 2L_u\|G\|e^{L(a_2-b_0)}(\varepsilon_1+\varepsilon_2) \\ &\vdots \\ &\le u\big(a_{p+1}, y(a_{p+1})\big) + 2L_u\|G\|e^{L(a_{p+1}-b_0)}\varepsilon^p \\ &= u\big(b, y(b)\big) + 2L_u\|G\|e^{L(b-a)}\varepsilon^p. \end{aligned}$$

By taking $p \to +\infty$, we have $\varepsilon^p \to 0$ and the desired result is obtained. □

Theorem 3.12 *Suppose u is a locally Lipschitz continuous subsolution of* (2.2a)–(2.2c). *Then u satisfies the sub-optimality, i.e. for any trajectory $y(\cdot) \in S_{[t,T]}(x)$, one has*

$$u(t,x) \le u\big(t+h, y(t+h)\big), \quad \forall h \in [0, T-t].$$

Proof Let $\mathcal{M}$ be a union of subdomains (manifolds or interfaces). Let $\bar{d}_{\mathcal{M}} \in \{0, \ldots, d\}$ be the minimal dimension of the subdomains in $\mathcal{M}$. We claim that for any $h \in [0, T-t]$ and any trajectory $y(\cdot)$ of (3.1) lying totally within $\mathcal{M}$, we have

$$u\big(a, y(a)\big) \le u\big(b, y(b)\big), \quad \text{for any } [a,b] \subseteq [t, t+h]. \tag{3.8}$$

The proof of (3.8) is based on an induction argument with regard to the minimal dimension $\bar{d}_{\mathcal{M}}$:

(HR) for $\tilde{d} \in \{1, \ldots, d\}$, suppose that for any $\mathcal{M}$ with $\bar{d}_{\mathcal{M}} \ge \tilde{d}$ and for any trajectory $y(\cdot)$ that lies within $\mathcal{M}$, (3.8) holds.

Step (1): Let us first check the case when $\tilde{d} = d$. In this case, $\bar{d}_{\mathcal{M}} = d$, then $\mathcal{M}$ is a union of d-manifolds which are disjoint by (H1). For any trajectory $y(\cdot)$ of (3.1) lying within $\mathcal{M}$, since $y(\cdot)$ is continuous, $y(\cdot)$ lies entirely in one of the d-manifolds, denoted by Ω_i. The subsolution property of u implies that

$$-\partial_t u(t,x) + \sup_{p \in F_i(x)} \big\{-p \cdot Du(t,x)\big\} \le 0$$

holds in the viscosity sense. Since the dynamics on Ω_i is F_i which is Lipschitz continuous, then by the classical theory u satisfies the sub-optimality along $y(\cdot)$ and (3.8) holds true.

Step (2): Now assume that **(HR)** is true for $\tilde{d} \in \{1, \ldots, d\}$, and let us prove that **(HR)** is true for $\tilde{d} - 1$. In this case, the minimal dimension of subdomains in $\mathcal{M}$ is $\bar{d}_{\mathcal{M}} = \tilde{d} - 1$, $\tilde{d} \in \{1, \ldots, d\}$. As an induction hypothesis, assume that for any trajectory that lies within a union of subdomains each with dimension greater than $\tilde{d}$, then (3.8) holds. Three cases can occur.

- If $\mathcal{M}$ contains only one subdomain, i.e. $\mathcal{M} = \mathcal{M}_k$ with dimension $\bar{d}_{\mathcal{M}}$ for some $k \in \{1, \dots, m + \ell\}$, then for any trajectory $y(\cdot)$ lying within $\mathcal{M}_k$, the subsolution property of u implies that u satisfies the sub-optimality along $y(\cdot)$ since the dynamics F_k^{new} is Lipschitz continuous on $\mathcal{M}_k$.
- If $\mathcal{M}$ contains more than one subdomain and $\mathcal{M}$ is connected, let $\mathcal{M}'_1, \dots, \mathcal{M}'_p$ be all the subdomains contained in $\mathcal{M}$ with dimension $\bar{d}_{\mathcal{M}}$. Then $\widetilde{\mathcal{M}} := \mathcal{M} \backslash (\cup_{k=1}^{p} \mathcal{M}'_k)$ is a union of subdomains with dimension greater than $\tilde{d}$. We note that $\mathcal{M}'_k \subseteq \overline{\widetilde{\mathcal{M}}}$ for each $k = 1, \dots, p$. Then by the induction hypothesis and Proposition 3.11, (3.8) holds true for any trajectory lying entirely within $\widetilde{\mathcal{M}} \cup \mathcal{M}'_1$. Then by applying Proposition 3.11 for $\widetilde{\mathcal{M}} \cup \mathcal{M}'_1$ and $\mathcal{M}'_2$, (3.8) holds true for any trajectory lying entirely within $\widetilde{\mathcal{M}} \cup \mathcal{M}'_1 \cup \mathcal{M}'_2$. We continue this process and finally we have (3.8) holds true for any trajectory lying entirely within $\mathcal{M} = \widetilde{\mathcal{M}} \bigcup (\cup_{k=1}^{p} \mathcal{M}'_k)$.
- If $\mathcal{M}$ is not connected, for any trajectory $y(\cdot)$ lying within $\mathcal{M}$, since $y(\cdot)$ is continuous, then $y(\cdot)$ lies within one connected component of $\mathcal{M}$. Then by the same argument as above, (3.8) holds true for $y(\cdot)$. And the induction step is complete.

Finally, to complete the proof of the theorem, we remark that for any trajectory $y(\cdot)$ of (3.1), by considering $\mathcal{M} = \mathbb{R}^d$ with $\bar{d}_{\mathcal{M}} = 0$, taking $a = t$, $b = t + h$ in (3.8) we have

$$u(t, x) \le u\big(t + h, y(t + h)\big),$$

which ends the proof. □

4 Proof of Theorem 2.4

Since v satisfies the super-optimality and sub-optimality, by Theorem 3.7 and Theorem 3.10 v is a viscosity solution of (2.2a)–(2.2c).

The uniqueness result is obtained by the following result of comparison principle.

Proposition 4.1 *Suppose that $u : [0, T] \times \mathbb{R}^d \to \mathbb{R}$ is Lipschitz continuous and $u(T, x) = \varphi(x)$ for any $x \in \mathbb{R}^d$.*

(i) *If u satisfies the super-optimality, then $v(t, x) \le u(t, x)$ for all $(t, x) \in [0, T] \times \mathbb{R}^d$;*
(ii) *If u satisfies the sub-optimality, then $v(t, x) \ge u(t, x)$ for all $(t, x) \in [0, T] \times \mathbb{R}^d$.*

Proof (i) For any $(t, x) \in [0, T] \times \mathbb{R}^d$, by the super-optimality of u, there exists a trajectory $\overline{y}_{t,x}$ such that

$$u(t, x) \ge u\big(T, \overline{y}_{t,x}(T)\big) = \varphi\big(\overline{y}_{t,x}(T)\big).$$

By the sub-optimality of v, we have

$$v(t,x) \le v\big(T, \overline{y}_{t,x}(T)\big) = \varphi\big(\overline{y}_{t,x}(T)\big).$$

Then we deduce that

$$v(t,x) \le u(t,x).$$

(ii) The proof is completed by the same argument by considering the super-optimality of v and the sub-optimality of u. □

5 Conclusion

In this paper, we have studied the system (1.1) in a general framework of the multi-domains with several interfaces. The existence and uniqueness result of the solution is studied under some junction conditions on the interfaces. The latter are derived by considering a control problem for which the value function satisfies the system (1.1) on each sub-domain Ω_i. The analysis of this value function indicates the information that should be considered on the interfaces in order to guarantee a continuous solution of the system.

Acknowledgements The authors are grateful to Peter Wolenski and Antonio Siconolfi for many helpful discussions.

Appendix A

Proof of Lemma 2.2 Note that although G is only usc on $\mathbb{R}^d$, G is Lipschitz continuous on Γ_j since G is the convexification of a finite group of Lipschitz continuous multifunctions on Γ_j. For any $x \in \Gamma_j$, there exists $\alpha > 0$ and a diffeomorphism $g \in C^{1,1}(\mathbb{R}^d)$ such that

$$\overline{B(x,\alpha) \cap \Gamma_j} = \big\{x : g(x) = 0\big\} \quad \text{and} \quad \nabla g(y) \neq 0, \quad \forall y \in \overline{B(x,\alpha)}.$$

We can take g as the signed distance function to Γ_j for instance. Then there exists $\beta > 0$ such that

$$\big\|\nabla g(y)\big\| \ge \beta, \quad \forall y \in B(x,\alpha) \cap \Gamma_j.$$

For any $w \in G(x) \cap \mathcal{T}_{\Gamma_j}(x)$, by the Lipschitz continuity of G there exists $v \in G(y)$ such that

$$\|w - v\| \le L_G \|x - y\|,$$

where L_G is the Lipschitz constant of $G(\cdot)$. Since $w \in \mathcal{T}_{\Gamma_j}(x)$, we have

$$w \cdot \nabla g(x) = 0.$$

Then

$$\big\| v \cdot \nabla g(x) \big\| = \big\| (v - w) \cdot \nabla g(x) \big\| \leq L_G \|\nabla g\| \|x - y\|.$$

Thus,

$$\begin{aligned} \big\| v \cdot \nabla g(y) \big\| &\leq \big\| v \cdot \nabla g(x) \big\| + \big\| v \cdot \big(\nabla g(y) - \nabla g(x)\big) \big\| \\ &\leq \big(L_G \|\nabla g\| + \|G\| L'_g\big) \|x - y\|, \end{aligned}$$

where L'_g is the Lipschitz constant of $\nabla g(\cdot)$. We consider the following three cases:

If $v \cdot \nabla g(y) = 0$, then $v \in \mathcal{T}_{\Gamma_j}(y)$ and we deduce that

$$w \in G(y) \cap \mathcal{T}_{\Gamma_j}(y) + L_G \|x - y\| B(0, 1).$$

If $v \cdot \nabla g(y) := -\gamma < 0$, let $p := \delta \nabla g(y) / \|\nabla g(y)\|$, then by (H4)(iv),

$$p \in G(y) \quad \text{and} \quad p \cdot \nabla g(y) := \tilde{\beta} \geq \delta \beta > 0.$$

We set

$$q := \frac{\tilde{\beta}}{\tilde{\beta} + \gamma} v + \frac{\gamma}{\tilde{\beta} + \gamma} p,$$

then $q \cdot \nabla g(y) = 0$, i.e. $q \in \mathcal{T}_{\Gamma_j}(y)$. And since $G(y)$ is convex, we have $q \in G(y)$. Then we obtain

$$\begin{aligned} \|w - q\| &\leq \|w - v\| + \|v - q\| \\ &\leq L_G \|x - y\| + \frac{\gamma}{\tilde{\beta} + \gamma} \|v - p\| \\ &\leq \left(L_G + \frac{L_G \|\nabla g\| + \|G\| L'_g}{\delta \beta} 2\|G\| \right) \|x - y\|, \end{aligned}$$

where we deduce that

$$w \in G(y) \cap \mathcal{T}_{\Gamma_j}(y) + L \|x - y\| B(0, 1), \tag{A.1}$$

with $L := L_G + 2\|G\|(L_G \|\nabla g\| + \|G\| L'_g)/\delta\beta$.

If $v \cdot \nabla g(y) > 0$, then by the same argument taking $p = -\delta \nabla g(y) / \|\nabla g(y)\|$, (A.1) holds true as well.

Finally, (A.1) implies the local Lipschitz continuity of $G(\cdot) \cap \mathcal{T}_{\Gamma_j}(\cdot)$ on Γ_j with the local constant L. □

Proof of Lemma 3.6 For $k = 1, \ldots, m + \ell$, we set

$$\begin{aligned} J_k^n &:= \big\{ t \in [0, t_n] : y(t) \in \mathcal{M}_k \big\}, \\ \mu_k^n &:= meas\big(J_k^n\big), \end{aligned}$$

$$\mathbb{K}(x) := \big\{k : \mu_k^n > 0, \forall n \in \mathbb{N}\big\}.$$

For each $k \in \mathbb{K}(x)$, we have $x \in \mathcal{M}_k$. Up to a subsequence, there exists $0 \leq \lambda_k \leq 1$ and $p_k \in \mathbb{R}^d$ so that

$$\frac{\mu_k^n}{t_n} \to \lambda_k, \qquad \sum_{k \in \mathbb{K}(x)} \lambda_k = 1, \qquad \frac{1}{\mu_k^n} \int_{J_k^n} \dot{y}(s)ds \to p_k$$

as $n \to +\infty$. By Proposition 3.4 and the Lipschitz continuity of F_k^{new}, we have

$$\begin{aligned}
p_k &= \lim_{n\to\infty} \frac{1}{\mu_k^n} \int_{J_k^n} \dot{y}(s)ds \\
&\in \lim_{n\to\infty} \frac{1}{\mu_k^n} \int_{J_k^n} F_k^{new}\big(y(s)\big)ds \\
&\subseteq \lim_{n\to\infty} \left[\frac{1}{\mu_k^n} \int_{J_k^n} F_k^{new}(x)ds + \frac{1}{\mu_k^n} \int_{J_k^n} L_k \big\| y(s) - x \big\| B(0,1)ds\right] \\
&\subseteq \lim_{n\to\infty} \left[F_k^{new}(x) + L_k \|F\| \left[\frac{1}{\mu_k^n} \int_{J_k^n} sds\right] B(0,1)\right] = F_k^{new}(x).
\end{aligned}$$

We then have

$$\begin{aligned}
p &= \lim_{n\to\infty} \frac{y(t_n) - x}{t_n} = \lim_{n\to\infty} \frac{1}{t_n} \int_0^{t_n} \dot{y}(s)ds \\
&= \sum_{k \in \mathbb{K}(x)} \lim_{n\to\infty} \frac{\mu_k^n}{t_n} \left[\frac{1}{\mu_k^n} \int_{J_k^n} \dot{y}(s)ds\right] \\
&= \sum_{k \in \mathbb{K}(x)} \lambda_k p_k \in \sum_{k \in \mathbb{K}(x)} \lambda_k F_k^{new}(x) \subseteq \text{co} \bigcup_{k \in \mathbb{K}(x)} F_k^{new}(x).
\end{aligned}$$

Now set $\mathcal{M} := \cup_{k \in \mathbb{K}(x)} \mathcal{M}_k$, and since $y(t_n) \in \mathcal{M}$ for all large n, we have $p \in \mathcal{T}_{\overline{\mathcal{M}}}(x)$. Then we obtain

$$p \in \left(\text{co} \bigcup_{k \in \mathbb{K}(x)} F_k^{new}(x)\right) \cap \mathcal{T}_{\overline{\mathcal{M}}}(x).$$

The fact that $F_k^{new}(z) \subseteq \mathcal{T}_{\mathcal{M}_k}(z)$ whenever $z \in \mathcal{M}_k$ implies

$$F_k^{new}(x) \cap \mathcal{T}_{\overline{\mathcal{M}}}(x) = F_k^{new}(x) \cap \mathcal{T}_{\overline{\mathcal{M}_k}}(x)$$

whenever $x \in \overline{\mathcal{M}_k}$. Hence

$$p \in \text{co} \bigcup_{k \in \mathbb{K}(x)} \left(F_k^{new}(x) \cap \mathcal{T}_{\overline{\mathcal{M}_k}}(x)\right) = \text{co}\ F^E(x). \qquad \square$$

Appendix B

We review the background in nonsmooth analysis required in our analysis. A closed set $\mathcal{C} \subseteq \mathbb{R}^d$ is called *proximally smooth* of radius $\delta > 0$ provided the distance function $d_{\mathcal{C}}(x) := \inf_{c \in \mathcal{C}} \|c - x\|$ is differentiable on the open neighborhood $\mathcal{C} + \delta B(0, 1)$ of $\mathcal{C}$. For any $c \in \mathcal{C}$, we denote the Clarke normal cone by $\mathcal{N}_{\mathcal{C}}(c)$. Recall the *tangent cone* $\mathcal{T}_{\mathcal{C}}(c)$ at $c \in \mathcal{C}$ is defined as

$$\mathcal{T}_{\mathcal{C}}(c) = \left\{ v : \liminf_{t \to 0^-} \frac{d_{\mathcal{C}}(c + tv)}{t} = 0 \right\},$$

and in the case of $\mathcal{C}$ proximally smooth, equals the Clarke tangent cone as the negative polar of $\mathcal{N}_{\mathcal{C}}(c)$:

$$v \in \mathcal{T}_{\mathcal{C}}(c) \quad \Longleftrightarrow \quad \langle \zeta, v \rangle \leq 0 \quad \forall \zeta \in \mathcal{N}_{\mathcal{C}}(c).$$

If $\mathcal{M}$ is an embedded C^2 manifold, $\mathcal{C} := \overline{\mathcal{M}}$, and $c \in \mathcal{M}$, then $\mathcal{T}_{\mathcal{C}}(c)$ agrees with the usual tangent *space* $\mathcal{T}_{\mathcal{M}}(c)$ to $\mathcal{M}$ at c from differential geometry (see [9, Proposition 1.9]). If in addition $\overline{\mathcal{M}}$ is proximally smooth, then for each $x \in \overline{\mathcal{M}}$, the tangent cone $\mathcal{T}_{\overline{\mathcal{M}}}(x)$ is closed and convex, and thus has a relative interior denoted by *r-int* $\mathcal{T}_{\overline{\mathcal{M}}}(x)$. Its relative boundary is defined as *r-bdry* $\mathcal{T}_{\overline{\mathcal{M}}}(x) := \mathcal{T}_{\overline{\mathcal{M}}}(x) \backslash$*r-int* $\mathcal{T}_{\overline{\mathcal{M}}}(x)$.

Another key assumption on the multi-domains is each domain being relatively wedged. A set $\mathcal{C} \subseteq \mathbb{R}^N$ is *wedged* (see [9, p.166]) if at every $x \in$ bdry $\mathcal{C}$, *int*$\mathcal{T}_{\overline{\mathcal{C}}} \neq \emptyset$. If $\mathcal{C} = \overline{\mathcal{M}}$ is the closure of an embedded manifold $\mathcal{M}$, then $\mathcal{C}$ relatively wedged means the dimension of *r-int* $\mathcal{T}_{\overline{\mathcal{M}}_k}(x)$ is equal to d_k.

The following result is [5, Lemma 3.1] and is the key geometrical ingredient that permits the construction of boundary trajectories of (DI).

Lemma B.1 *If $x \in \overline{\mathcal{M}_k} \backslash \mathcal{M}_k$ and $v \in$ r-bdry $\mathcal{T}_{\overline{\mathcal{M}_k}}(x)$, then there exists an index j for which $\mathcal{M}_j \subseteq \overline{\mathcal{M}_j}$, $x \in \overline{\mathcal{M}}_j$, and $v \in \mathcal{T}_{\overline{\mathcal{M}_k}}(x)$. Of course in this case, one has $d_j < d_k$.*

Proof See [5, Lemma 3.1]. □

We finally give a sketch of the proof for Lemma 3.8.

Proof A key fact is that for any $p \in G(x) \cap$ *r-int* $\mathcal{T}_{\overline{\mathcal{M}}_k}(x)$, there exist $\tau > 0$ and a C^1 trajectory $y(\cdot) : [t, \tau] \to \mathcal{M}_k \cup \{x\}$ so that

$$y(t) = x \quad \text{and} \quad \dot{y}(t) = p. \tag{B.1}$$

Let k be such that $x \in \overline{\mathcal{M}_k}$ and $p \in G(x) \cap \mathcal{T}_{\overline{\mathcal{M}}_k}(x)$. If $p \in$ *r-int* $\mathcal{T}_{\overline{\mathcal{M}}_k}(x)$, then the result follows by the key fact (B.1). If $p \notin$ *r-int* $\mathcal{T}_{\overline{\mathcal{M}}_k}(x)$, then $p \in$ *r-bdry* $\mathcal{T}_{\overline{\mathcal{M}}_k}(x)$ and hence by Lemma B.1, there exists another subdomain $\mathcal{M}_j \subseteq \overline{\mathcal{M}_k}$ with $x \in \overline{\mathcal{M}_j}$

and $p \in \mathcal{T}_{\overline{\mathcal{M}}_j}(x)$. If $p \in$ *r-int* $\mathcal{T}_{\overline{\mathcal{M}}_j}(x)$, then the result follows from (B.1), otherwise the argument just given can be repeated with k replaced by j. The process must eventually terminate since the dimension is decreasing at each step. □

References

1. Y. Achdou, F. Camilli, A. Cutri, N. Tchou, Hamilton–Jacobi equations on networks. Nonlinear Differ. Equ. Appl. (2012). doi:10.1007/s00030-012-0158-1
2. J.-P. Aubin, A. Cellina, *Differential Inclusions*. Comprehensive Studies in Mathematics, vol. 264 (Springer, Berlin, 1984)
3. M. Bardi, I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations*. Systems and Control: Foundations and Applications (Birkhäuser, Boston, 1997)
4. G. Barles, A. Briani, E. Chasseigne, A Bellman approach for two-domains optimal control problems in $\mathbb{R}^N$. To appear in ESAIM Control Optim. Calc. Var. (2012)
5. R.C. Barnard, P.R. Wolenski, Flow invariance on stratified domains. Set-Valued Var. Anal. (2013). doi:10.1007/s11228-013-0230-y
6. A. Bressan, Y. Hong, Optimal control problems on stratified domains. Netw. Heterog. Media **2**(2), 313–331 (2007)
7. F. Camilli, A. Siconolfi, Hamilton–Jacobi equations with measurable dependence on the state variable. Adv. Differ. Equ. **8**(6), 733–768 (2003)
8. F.H. Clarke, *Optimization and Nonsmooth Analysis* (Society for Industrial and Applied Mathematics, Philadelphia, 1990)
9. F.H. Clarke, Yu.S. Ledyaev, R.J. Stern, P.R. Wolenski, *Nonsmooth Analysis and Control Theory* (Springer, Berlin, 1998)
10. G. Dal Maso, H. Frankowska, Value function for Bolza problem with discontinuous Lagrangian and Hamilton–Jacobi inequalities. ESAIM Control Optim. Calc. Var. **5**, 369–394 (2000)
11. A.F. Filippov, *Differential Equations with Discontinuous Right-Hand Sides* (Kluwer Academic, Norwell, 1988)
12. H. Frankowska, Lower semicontinuous solutions of Hamilton–Jacobi–Bellman equations. SIAM J. Control Optim. **31**(1), 257–272 (1993)
13. C. Imbert, R. Monneau, H. Zidani, A Hamilton–Jacobi approach to junction problems and application to traffic flows. ESAIM Control Optim. Calc. Var. **e-first** (2011)
14. H. Ishii, A boundary value problem of the Dirichlet type for Hamilton–Jacobi equations. Ann. Sc. Norm. Super. Pisa, Cl. Sci. **16**(1), 105–135 (1989). (eng)
15. Z. Rao, A. Siconolfi, H. Zidani, Transmission conditions on interfaces for Hamilton–Jacobi–Bellman equations (2013 submitted)
16. P. Soravia, Boundary value problems for Hamilton–Jacobi equations with discontinuous Lagrangian. Indiana Univ. Math. J. **51**(2), 451–477 (2002)

Gradient Computation for Model Calibration with Pointwise Observations

Ekkehard W. Sachs and Matthias Schu

Abstract Mathematical models for option pricing often result in partial differential equations of parabolic type. The calibration of these models leads to an optimization problem with PDE constraints and usually pointwise observations in the objective function. Thus, the adjoint equation of this problem involves Dirac delta functions and needs a special treatment from a numerical point of view. We show by means of numerical results that also the order of discretizing and optimizing plays an important role.

Keywords Option pricing models · Pointwise observations · Adjoint · Optimize · Discretize

1 Introduction

The pricing of financial derivatives, e.g., European call options, plays an important role in many finance applications. Thus, there is a vast literature on the theory and numerics of option pricing, see, e.g., the following references for an introduction, [8, 17] and [31].

The most famous option pricing model is the Black–Scholes model introduced in [6] and [22]. Here, the price can be calculated via the solution of a partial differential equation.

A more advanced model considered in this article is based on so-called jump-diffusion processes and leads to partial integro-differential equations (PIDE). Since in general no closed-form solution is available, they have to be solved numerically, leading to dense systems of equations that need a special treatment, see [2, 3, 27] and [25].

Support acknowledged by DFG Special Priority Program 1253 ‘Optimization with Partial Differential Equations’.

E.W. Sachs (✉) · M. Schu
Fachbereich IV, Abteilung Mathematik, Universität Trier, 54286 Trier, Germany
e-mail: sachs@uni-trier.de

M. Schu
e-mail: schu@uni-trier.de

K. Bredies et al. (eds.), *Control and Optimization with PDE Constraints*,
International Series of Numerical Mathematics 164,
DOI 10.1007/978-3-0348-0631-2_7, © Springer Basel 2013

An important aspect regarding option pricing models is the proper choice of the model parameters. From a mathematical point of view, this task leads to a constrained calibration problem, where model prices are compared with given market prices in a least-square setting subject to the PIDE. In practice, market prices are only available for certain options on a particular underlying. In the objective function of the calibration problem, these pointwise observations lead to the use of Dirac delta functions in the gradient formulation using the adjoint calculus.

Establishing optimality conditions for the optimization problem, we have to make a decision, whether to optimize or discretize first. This has been the topic of papers for several decades starting in the early days of numerical analysis for optimal control problems.

In contrary to the opinion of part of the research community, this article documents that there are cases when optimizing first leads to significantly better results in terms of gradient approximations and a more exact calculation of the adjoint equation. The reason for this observation lies in the fact that formulating the optimality conditions or calculating the gradient of the corresponding unconstrained problem involves the adjoint equation. The pointwise observations in the calibration problem show up in the adjoint equation in terms of Dirac delta functions. This could lead problematic behavior of the numerical scheme and needs a special numerical treatment. In contrast, optimizing first and choosing an appropriate numerical scheme afterward gives consistently better results.

In Sect. 2 of this paper, we give a short introduction into the PIDE which occurs in the pricing of options where jump-diffusion processes are the theoretical background model. In the second part of this section, we formulate the weak solution of the PIDE for a localized version. We take a special look at the initial condition which is not smooth and where the numerical results can be improved by the use of Rannacher smoothing techniques. This observation will become important at a later stage of the paper.

Section 3 contains the formulation of a calibration problem using a Dupire-like forward PIDE. The following Sect. 4 contains the definition of an adjoint PIDE and outlines the first-optimize-then-discretize approach. In Sect. 5 we consider the first-discretize-then-optimize calculus. Both approaches are considered for the computation of gradients, since this is an essential ingredient of any optimization algorithm.

The final section contains numerical results and shows that the adjoint equation contains very nonsmooth initial data. Without a smoothing technique like Rannacher smoothing as illustrated in Sect. 2, the numerical results are useless. This shows the importance of this technique in the framework of point observations.

The Rannacher smoothing technique has been considered for option pricing in connection with adaptive schemes in [14]. A weak solution concept using pointwise observations in parabolic differential equations has recently also been addressed in [15].

2 Option Pricing Models

The most famous option pricing model is the Black–Scholes model introduced in [6] and [22]. Here, the price can be calculated via the solution of a partial differential equation.

Since practitioners are aware of several shortcomings of this model (cf. [3, 9, 28] and [11]), more advanced approaches have been introduced that can be grouped into three main ideas (cf. [3]).

[12] and [11] proposed the so-called 'local volatility models'. Here, the constant volatility of the Black–Scholes model is replaced by a deterministic function of the stock price S and the time t, $\sigma(t, S)$.

Although there is no closed-form solution available as in the Black–Scholes case, the model can be handled in a standard way from a numerical point of view. Since $\sigma(t, S)$ is a function, we can fit the model precisely to many quoted call prices. Beside these advantages, there are also some drawbacks. Especially, the fitting to typical skews of the implied volatility for short-term calls requires an unrealistic heavily twisting of the local volatility surface.

The second Black–Scholes generalization to mention is the 'stochastic volatility model'. There are different concrete approaches by, e.g., [18, 30] and the most famous by Heston (cf. [16]). The latter models the dynamics of the stock price by a Brownian motion like the Black–Scholes model, but in addition, the volatility as the driving force of the call price is modeled by a stochastic process, namely an Ornstein–Uhlenbeck process. There are parameter combinations, where the implied volatilities of the corresponding Heston prices show the typical skew or smile, that is observed in market data. But often the correlation between stock price and volatility has to be chosen unrealistically high to get the desired result. On the other hand, an important advantage of the Heston model is the existence of a closed-form solution.

The third approach requires the introduction of a new stochastic process, since the uncertainty is no longer modeled solely by a continuous Brownian motion. [23] suggested to add random jumps as an additional source of uncertainty in order to model also rare large market movements. This more general stochastic process compared to the Brownian motion is called Lévy process.

The option pricing models based on general exponential Lévy processes can be divided into two categories. One is called 'jump-diffusion models'. Here, as proposed by [23], jumps are added to the Brownian motion to model large movements of the asset. The second type are so-called 'infinite activity models', where the Brownian motion is omitted and more or less replaced by an infinite number of small jumps. [4, 7] can be named as references for models of the last-mentioned type. In this paper, we focus on jump-diffusion models driven by the following stochastic differential equation

$$dS_t = \mu S_{t\_} dt + \sigma S_{t\_} dW_t + S_{t\_} d\left(\sum_{j=1}^{N_t} \left(e^{Y_j} - 1\right)\right), \tag{2.1}$$

where $\sigma > 0$. The first two parts on the right-hand side are equal to the Black–Scholes model, a drift term and a Brownian motion. However, a third term is added,

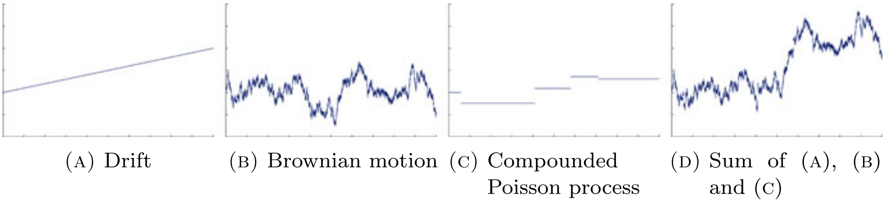


Fig. 1 Composition of a typical path of a jump diffusion process $X_t = \mu t + \sigma W_t + \sum_{j=1}^{N_t} Y_j$ used to model the log-price

where jumps enter the process by a compounded Poisson process. Figure 1 shows how a typical path of such a process is composed of these three terms.

Jump-diffusion models differ by the distribution of the jump sizes Y_j. There are two popular examples.

Example 2.1

1. Merton [23]: Y_i are normally distributed with the well-known density function

$$f^M(y) = \frac{1}{\sqrt{2\pi}\sigma_J} \exp\left\{-\frac{(y-\mu_J)^2}{2\sigma_J^2}\right\}.$$

2. Kou [20]: Y_i has an asymmetric double exponential distribution with density

$$f^K(y) = p\lambda^+ e^{-\lambda^+ y}\mathbb{1}_{\{y\geq 0\}} + (1-p)\lambda^- e^{\lambda^- y}\mathbb{1}_{\{y<0\}}.$$

The main advantage of these new models for the stock price dynamics is that a skew in the implied volatility, especially for short-term options, can be produced quite easily by setting the mean jump size to a negative value. This eliminates one of the main weaknesses of the approaches mentioned before. On the other hand, the models by Merton and Kou as well as the stochastic volatility models contain only a few parameters that can be calibrated to market prices. Thus, given many market prices, the calibration problem could be significantly underdetermined and errors between market and model prices might be too large. Furthermore, jump-diffusion models are known to be difficult to handle from a numerical point of view, although at least for the Merton model (see Example 2.1), there exists a semi-analytical solution in terms of a series of Black–Scholes prices (cf. [23]).

Remark 2.2 Given the Merton jump-diffusion model (see (2.1) and Example 2.1) with volatility $\sigma > 0$, jump intensity λ, mean jump size μ_J and volatility of the jump size σ_J. Then for a given maturity T, strike price K, interest rate r and current stock price S_0, the price of a European call option is given by

$$C^M = \sum_{n=0}^{\infty} \frac{e^{-\bar{\lambda}T}(\bar{\lambda}T)^n}{n!} C^{BS}(0, S_0;\ \bar{r}, \bar{\sigma}),$$

where $\bar{\lambda} = \lambda(1 + \mu_J)$, $\bar{r} = r - \lambda\mu_J + n\ln(1 + \mu_J)/T$, $\bar{\sigma}^2 = \sigma^2 + n\sigma_J^2/T$ and $C^{BS}(0, S_0;\ \bar{r}, \bar{\sigma})$ is the Black–Scholes price with interest rate $\bar{r}$ and volatility $\bar{\sigma}$.

Since all three generalizations of the Black–Scholes model described above still have some weaknesses, there is a vast literature on combinations of the different approaches. [5] combined stochastic volatility with jump-diffusion. [19] and [21] proposed a stochastic local volatility model, where [21] also includes jumps, i.e. a combination of all three approaches. A jump-diffusion model with local volatility was suggested by [3]. Since it is suitable to produce typical volatility skews with the jump part and to fit prices closely with the local volatility function, we will use this model in the following. Hence, the development of the stock price in time is modeled by the following stochastic differential equation:

$$dS_t = \mu S_{t\_}dt + \sigma(t, S_{t\_})S_{t\_}dW_t + S_{t\_}d\left(\sum_{j=1}^{N_t}\left(e^{Y_i} - 1\right)\right). \tag{2.2}$$

In the following section we take a closer look at the fair price of a European call option based on this stochastic model.

2.1 Partial Integro-Differential Equations

Due to the fact, that especially in case of local volatility functions, there is no analytical solution available, there is a need for a numerical solution of the problem. There are two main approaches which could be followed: The Monte Carlo simulation of the stochastic differential equation or the transformation of the SDE into a partial differential equation that can be solved numerically. We follow the latter approach.

We consider a Dupire-like version of the PIDE, i.e. the strike price and maturity of an option are treated as variables in the PIDE model. This is the suitable formulation for a model calibration. For today's price of a European call option on a given underlying asset the following holds (cf. [3] or [1]):

Theorem 2.3 *Given a jump-diffusion model as in* (2.2), *then today's price of a European call option $C(T, K)$ with fixed current underlying price S_0, maturity T and strike price K can be calculated via the partial integro-differential equation*

$$\begin{aligned} &C_T(T, K) - \frac{1}{2}\sigma^2(T, K)K^2C_{KK}(T, K) + r(T)KC_K(T, K) \\ &\quad - \lambda\int_{-\infty}^{+\infty}\left(e^y\left(C\left(T, Ke^{-y}\right) - C(T, K)\right) + K\left(e^y - 1\right)C_K(T, K)\right)f(y)\,dy = 0 \end{aligned} \tag{2.3}$$

$$(T, K) \in [0, T_{max}) \times (0, \infty)$$

with initial condition $C(0, K) = \max\{S_0 - K, 0\}$, $K \in (0, \infty)$.

2.2 Numerical Solution of the PIDE

We now briefly discuss the numerical solution of the PIDE introduced above. For the discretization in space, we choose a finite element approach; a finite difference approach leads to similar results. We set

$$\zeta = \zeta(f) := \int_{-\infty}^{+\infty} \left(e^{y} - 1\right) f(y)\, dy$$

and apply a variable transformation to the 'log-moneyness' $x = \ln(K/S_0)$, before we formulate the variational formulation of problem (2.3). Note that the variable transformation above leads to a problem defined now on the domain $[0, T_{max}] \times (-\infty, +\infty)$. To solve this numerically, the problem has to be localized to a domain $[0, T_{max}] \times [\underline{x}, \overline{x}]$ by introducing appropriate boundary conditions. We set $H := L^2(\underline{x}, \overline{x})$, $V := H_0^1(\underline{x}, \overline{x})$, $W([a, b], V) := \{u : u \in L^2((a, b), V), u' \in L^2((a, b), V^*)\}$ with $a < b \in \mathbb{R}$ (cf. [10]), and formulate a weak solution of the PIDE as follows.

Definition 2.4 A localized variational formulation of the PIDE (2.3) consists of finding $y \in W([0, T_{max}], V)$ such that for all $T \in (0, T_{max}]$

$$\frac{d}{dT}\big\langle y(T,\cdot), w(\cdot)\big\rangle_H + a\big(T; y(T,\cdot), w(\cdot)\big) = L\big(T;\ w(\cdot)\big) \quad \forall w \in V \tag{2.4}$$

holds with initial condition

$$\big\langle y(0,\cdot), w(\cdot)\big\rangle_H = \big\langle \hat{y}(\cdot), w(\cdot)\big\rangle_H \quad \forall w \in V. \tag{2.5}$$

The time-dependent bilinear form $a(T; \cdot, \cdot) : V \times V \to \mathbb{R}$ is given by

$$\begin{aligned} a(T;\ v, w) := &\int_{\underline{x}}^{\overline{x}} \frac{\sigma^2(T, S_0 e^x)}{2} v'(x) w'(x) dx \\ &+ \int_{\underline{x}}^{\overline{x}} \left(r(T) + \frac{\sigma^2(T, S_0 e^x)}{2} - \lambda\zeta + \frac{(\sigma^2(T, S_0 e^x))_x}{2} \right) v'(x) w(x) dx \\ &+ \int_{\underline{x}}^{\overline{x}} \lambda(1+\zeta) v(x) w(x) dx - \lambda \int_{\underline{x}}^{\overline{x}} \int_{\mathbb{R}} v(x-y) w(x) e^y f(y) dy dx, \end{aligned} \tag{2.6}$$

$L(T;\ \cdot) : V \to \mathbb{R}$ is a time-dependent linear functional containing the boundary conditions $y^b(T, x)$, and $\hat{y}(x) := \max\{1 - e^x, 0\} - y^b(0, x)$ is an adjusted initial condition with zero boundary.

For a more detailed discussion on the boundary conditions $y^b(T, x)$, the localization and the variational formulation above, we refer to [25]. In the numerical results below, we will use linear splines to approximate the Hilbert space V in a Galerkin approach.

Before we proceed with the time discretization, we notice that (2.4) and (2.5) can be rewritten in operator form in the sense of $L^2(V^*)$. This has some advantages in terms of a more simple notation.

Remark 2.5 There exist unique operators $A(t) \in L(V, V^*)$, $l(t) \in V^*$ for all $t \in (0, T_{max}]$ and $\hat{y} \in H$ such that (2.4) and (2.5) can be rewritten as

$$\dot{y}(t) + A(t)y(t) = l(t) \quad \forall t \in (0, T_{max}], \tag{2.7}$$

$$y(0) = \hat{y}, \tag{2.8}$$

in the sense of $L^2(V^*)$.

In the following we address the time discretization. It is well-known that implicit methods lead to dense linear systems of equations in the case of partial integro-differential equations. However, the special structure of the double integral term in (2.6) can be used to achieve nearly linear complexity in space even for implicit methods by using a preconditioned GMRES algorithm (cf. [25] for details).

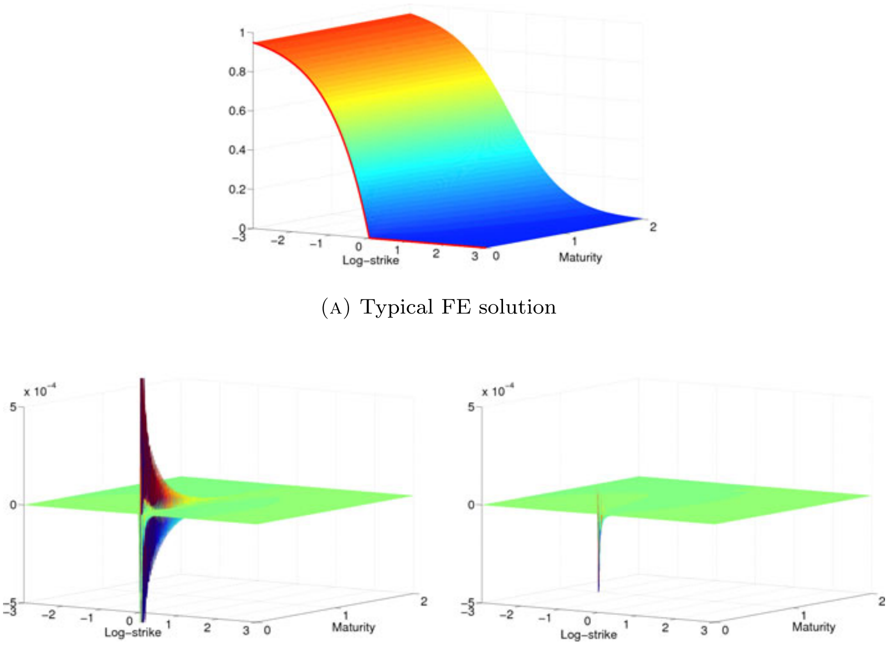
It is well-known that the problem above is stiff, so an unstable forward Euler method would be severely restricted by the CFL condition. Among the stable schemes Crank–Nicolson would be the method of choice regarding the error of the time discretization since the central difference quotient is of order $\mathcal{O}(\Delta t^2)$. But although it is A-stable, non-smooth initial conditions often lead to oscillations. [24] proposed to use two half-time steps of the strongly A-stable backward Euler scheme to smoothen the initial condition and then to proceed with Crank–Nicolson to preserve second order convergence. In [13], four backward Euler full- or four quarter-timesteps are named as an alternative to the original approach. This Rannacher approach will be of special interest in the subsequent sections, when we solve the adjoint equation.

We present some numerical results for Merton's jump diffusion model to illustrate the behavior of the Rannacher time-stepping. The model constants are set as follows:

$$\underline{x} = -5, \quad \overline{x} = 5, \quad T_{max} = 2y, \quad r \equiv 3\ \%,$$

$$\sigma \equiv 30\ \%, \quad \lambda = 100\ \%, \quad \mu_J = 0\ \%, \quad \sigma_J = 50\ \%, \quad \Delta x = 0.005, \quad \Delta T = 0.01.$$

First notice the non-smooth initial condition illustrated in Fig. 2a with a typical numerical solution of problem (2.4), (2.5). Figure 2b shows the error between the closed-form solution (cf. Remark 2.2) and an ordinary Crank–Nicolson time discretization. As expected, there are oscillations due to the non-smooth initial condition. If we replace the first Crank–Nicolson step by four Rannacher quartersteps, i.e. applying the Rannacher smoothing, the result illustrated in Fig. 2c is a far better approximation, especially at $x = 0$, a region, which is important for practical applications. Tables 1 and 2 support the observations of Fig. 2. They show the error between a numerical solution with Crank–Nicolson and Rannacher time-stepping, respectively, and the closed-form solution for Merton's model. The $L^\infty(\Omega)$- and



(A) Typical FE solution

(B) Error of Crank–Nicolson method

(C) Error with Rannacher smoothing

Fig. 2 A typical FE solution and the error between FE solution and closed-form solution for the Merton model ($\Delta x = 0.005$, $\Delta T = 0.01$)

Table 1 $L^2(\Omega)$- resp. $L^\infty(\Omega)$-error (for $T = 1$ and $T = 2$) between finite element solution with Crank–Nicolson and closed-form solution for different time step sizes ΔT and fixed Δx

Discretization		$L^\infty(\Omega)$-error				$L^2(\Omega)$-error			
Δx	ΔT	$T = 1$	Ratio	$T = 2$	Ratio	$T = 1$	Ratio	$T = 2$	Ratio
0.00125	0.08	2.20e-3		1.51e-3		1.73e-4		1.06e-4	
	0.04	1.01e-3	2.2	6.51e-4	2.3	6.26e-5	2.8	3.73e-5	2.8
	0.02	4.13e-4	2.4	2.50e-4	2.6	2.21e-5	2.8	1.31e-5	2.8
	0.01	1.41e-4	2.9	8.45e-5	3.0	7.63e-6	2.9	4.24e-6	3.1

$L^2(\Omega)$-error is evaluated at the time instances $T = 1$ and $T = 2$ on the domain $\Omega = [-3, 3]$. The columns captioned by ‘ratio’ show the factor of decrease in the error when the number of discretization steps in time is doubled. Note that the spatial discretization is chosen very fine to guarantee that the error is mainly driven by the time discretization error.

It is observable that the Crank–Nicolson method does not show a quadratic convergence, what would be indicated by a ratio of 4. However, this ratio is visible for the Rannacher smoothing in Table 2. After having sketched the numerical solution

Table 2 $L^2(\Omega)$- resp. $L^\infty(\Omega)$-error (for $T=1$ and $T=2$) between finite element solution with Rannacher smoothing and closed-form solution for different time step sizes ΔT and fixed Δx

Discretization		$L^\infty(\Omega)$-error				$L^2(\Omega)$-error			
Δx	ΔT	$T=1$	Ratio	$T=2$	Ratio	$T=1$	Ratio	$T=2$	Ratio
0.00125	0.08	2.39e-5		1.11e-5		2.48e-5		1.37e-5	
	0.04	6.41e-6	3.7	2.77e-6	4.0	6.41e-6	3.9	3.42e-6	4.0
	0.02	1.63e-6	3.9	6.79e-7	4.1	1.61e-6	4.0	8.58e-7	4.0
	0.01	4.24e-7	3.8	1.59e-7	4.3	4.14e-7	3.9	2.21e-7	3.9

of the option pricing problem and some of the numerical challenges, we turn to the calibration of the model parameters.

3 The Calibration Problem

A proper choice of the parameters of the option pricing model is quite important and shows the efficiency of the selected model. To guarantee that the parameters include the latest market information, the parameters are chosen in such way that the according model prices $C(T_i, K_i)$ for options on different strikes K_i and maturities T_i, $i=1,\dots,M$, fit the current market prices C_i^M. This model calibration is performed for frequently traded options as European or American call and put options. These parameters can then be used to calculate prices for new or more complex options. The typical calibration problem is formulated in a least-square sense.

Definition 3.1 The calibration problem consists of finding parameters $\sigma(\cdot,\cdot)$, λ and $f(\cdot)$, that solve the following minimization problem

$$\min_{C,\sigma,\lambda,f} J(C,\sigma,\lambda,f) := \frac{1}{2}\sum_{i=1}^{M}\big(C(T_i,K_i) - C_i^M\big)^2 \tag{3.1}$$

$$\begin{aligned} s.t.\quad & C_T - \frac{1}{2}\sigma^2(T,K)K^2C_{KK} + \big(r(T) - \lambda\zeta\big)KC_K + \lambda(1+\zeta)C \\ & - \lambda\int_{-\infty}^{+\infty} C\big(T, Ke^{-y}\big)e^y f(y)\,dy = 0, \quad (T,K) \in [0,T_{max}) \times (0,\infty), \end{aligned}$$

$$\text{with initial condition } C(0,K) = \max\{S_0 - K, 0\}, \qquad K \in (0,\infty),$$

where C_i^M are market prices for European call options with strike K_i and maturity T_i, $i=1,\dots,M$.

A proper regularization term might be added to the objective function. Since we are mainly interested in the gradient error, a regularization is not our main concern.

However, for the sake of completeness, we will introduce a simple Tikhonov regularization term below.

Thus, the calibration problem is a PIDE constrained optimization problem. Note that for one function evaluation of J, the PIDE constraint has to be solved only once. This is due to the use of the Dupire-like forward equation.

Before we look at the numerical solution of the calibration problem, we rewrite it in a more abstract way, in which the PIDE is replaced by its weak formulation according to Remark 2.5. This has some advantages especially in terms of a more simple notation.

For this purpose we denote by V and H two real, separable Hilbert spaces with $V \hookrightarrow H = H^* \hookrightarrow V^*$, and by $\mathcal{U}$, a closed, convex subspace of a Hilbert space, the space of control variables. $A(u;t) \in \mathcal{L}(V, V^*)$ is the elliptic operator, which is Fréchet-differentiable with respect to the control variable $u \in \mathcal{U}$ with a Fréchet derivative $A'(u;t)(\cdot) : \mathcal{U} \to \mathcal{L}(V, V^*)$, $t \in [0,T]$. The right-hand side of the equation is then denoted by $l(u;\cdot) \in L^2(V^*)$ for all $u \in \mathcal{U}$ with a Fréchet derivative $l'(u;t)(\cdot) : \mathcal{U} \to V^*$, $t \in [0,T]$.

Market data $d_i \in \mathcal{H}$ are available at certain maturities $\hat{t}_i$, $i = 1, \dots, D$, where $\mathcal{H}$ is a Hilbert space with $\mathcal{H}^* = \mathcal{H}$. For instance, $\mathcal{H} = \mathbb{R}^5$, if we have data available for five strike prices at maturity $\hat{t}_i$. We assume $\hat{t}_i < \hat{t}_j$ for $i < j$ and define $\hat{t}_0 = 0$ and $\hat{t}_D = T$. $C \in \mathcal{L}(H, \mathcal{H})$ denotes the observation operator.

Given this setting we can now rephrase the optimal control problem in (3.1) in an abstract form.

Definition 3.2 For given market data d_i at $\hat{t}_i$ $(i = 1, \dots, D)$ and $\alpha > 0$, find solutions $y \in W([0,T], V)$ and $u \in \mathcal{U}$, which solve the optimization problem

$$\min_{y \in W, u \in \mathcal{U}} J(y,u) := \min_{y \in W, u \in \mathcal{U}} \frac{1}{2} \sum_{i=1}^{D} \|Cy(\hat{t}_i) - d_i\|_{\mathcal{H}}^2 + \frac{\alpha}{2}\|u\|^2 \tag{3.2}$$

$$\begin{aligned} s.t. \quad & \dot{y}(t) + A(u;t)y(t) - l(u;t) = 0, \quad t \in (0,T], \\ & y(0) = \hat{y}. \end{aligned}$$

Remark 3.3 In (3.1) the objective function involves pointwise observations. Assuming $\mathcal{H} = \mathbb{R}^1$, i.e. that market data is given for only one strike price $\hat{x}$ at maturity $\hat{t}_i$, then the observation operator C would include a Dirac delta function, which is known to be not L^2-integrable. To avoid the involvement of distribution theory, we address the numerical approximation of C already at this stage and interpret C as an L^2-approximation $\delta_{\hat{x}}^{\Delta x}$ of the Dirac delta function. [29] proposed and analyzed a concrete representation that is equivalent to $\delta_{\hat{x}}$ on a discretized finite element space $\mathcal{V}^n$ in the sense that $\langle \delta_{\hat{x}}^{\Delta x}, v \rangle_{L^2} = \delta_{\hat{x}}(v)$, $v \in \mathcal{V}^n$. We have in this special case

$$\begin{aligned} Cv &= \langle \delta_{\hat{x}}^{\Delta x}, v \rangle_{L^2}, \quad v \in L^2, \\ C^* z &= \delta_{\hat{x}}^{\Delta x} z, \quad z \in \mathbb{R}. \end{aligned}$$

Note that—although it is not indicated in the following—C depends on Δx in this case.

The bilinear form which defines $A(u;t)$ is coercive and continuous, thus, it is clear that for every $u \in \mathcal{U}$, the parabolic constraint admits a unique solution $y(u;\cdot) \in W([0,T],V)$. Note that this also holds true for the option pricing problem in Definition 2.4 under some specific assumptions, e.g., $\sigma(T,x) \geq \sigma_{min} > 0$ (see [26] for details). Hence, the problem specified in Definition 3.2 can be written as an unconstrained optimization problem, in literature also known as 'reduced problem'.

Remark 3.4 The problem specified in Definition 3.2 can be written as unconstrained optimization problem:

$$\min_{u\in\mathcal{U}} f(u) := \min_{u\in\mathcal{U}} J\big(y(u),u\big). \tag{3.3}$$

If we want to derive optimality conditions of (3.2) or a gradient representation of the unconstrained problem (3.3), we have to decide whether we discretize or optimize first. It turns out that this decision is of importance in our application from a numerical point of view. We will briefly discuss the two approaches.

4 First Optimize

In order to derive a gradient representation, $\nabla f(u)$, for the unconstrained problem, which is needed in an optimization algorithm, we first define the Lagrange function for problem (3.2).

Given Lagrange multipliers p^i, $i = 1,\ldots,D$, we define

$$\begin{aligned}\mathcal{L}\big(y,u,p^1,\ldots,p^D\big) &:= J(y,u)\\ &\quad + \sum_{i=1}^{D}\int_{\hat{t}_{i-1}}^{\hat{t}_i} \big\langle p^i(t), \dot{y}(t) + A(u;t)y(t) - l(u;t)\big\rangle_{V,V^*} dt,\end{aligned} \tag{4.1}$$

and are now able to derive heuristically the corresponding optimality conditions. We obtain the state equation

$$\begin{aligned}&\dot{y}(t) + A(u;t)y(t) - l(u;t) = 0, \qquad t \in (0,T]\\ &y(0) = \hat{y},\end{aligned} \tag{4.2}$$

where the initial condition $y(0) = \hat{y}$ is given explicitly since we did not introduce an additional Lagrange multiplier in (4.1); for $i = D$, the adjoint equation can be specified as

$$\dot{p}^D(t) - A^*(u;t)p^D(t) = 0, \quad t \in [\hat{t}_{D-1}, T), \tag{4.3}$$

$$p^D(T) = -C^*\big(Cy(T) - d_D\big) \tag{4.4}$$

and for $i = 1, \dots, D-1$

$$\dot{p}^i(t) - A^*(u;t)p^i(t) = 0, \quad t \in [\hat{t}_{i-1}, \hat{t}_i), \tag{4.5}$$

$$p^i(\hat{t}_i) = -C^*\big(Cy(\hat{t}_i) - d_i\big) + p^{i+1}(\hat{t}_i). \tag{4.6}$$

Note that in (4.6) $p^{i+1}(\hat{t}_i)$ is known since we start solving the adjoint equations backwards at $t = T$, i.e. we first solve (4.3) backwards with end condition (4.4). It can be shown easily that $p^i \in W([\hat{t}_{i-1}, \hat{t}_i], V)$, $i = 1, \dots, D$, if $C \in \mathcal{L}(H, \mathcal{H})$.

The partial derivative of the Lagrange function with respect to the control u along a direction δu leads to:

$$\sum_{i=1}^{D} \int_{\hat{t}_{i-1}}^{\hat{t}_i} \big\langle p^i(t), A'(u;t)\delta u\ y(t) - l'(u;t)\delta u \big\rangle_{V,V^*} dt + \alpha\langle u, \delta u\rangle \geq 0. \tag{4.7}$$

To show formally that (4.7) is the gradient of the unconstrained optimization problem, we introduce the sensitivity $z(t) = \frac{\partial y(u;t)}{\partial u}\delta u$. The following result is true:

Lemma 4.1 *Given $u \in \mathcal{U}$, the corresponding solution $y(u;\cdot) \in W([0,T], V)$ and a direction δu. Further let z be the unique solution of*

$$\begin{aligned} &\dot{z}(t) + A(u;t)z(t) + A'(u;t)\delta u\ y(u;t) - l'(u;t)\delta u = 0, \\ &z(0) = 0. \end{aligned} \tag{4.8}$$

Then $z \in W([0,T], V)$ is the Fréchet derivative of $y(u;\cdot)$ with respect to u along direction δu.

Theorem 4.2 *The derivative of $f(u)$ (defined in (3.3)) along a feasible direction δu is given by*

$$f'(u)\delta u = \sum_{i=1}^{D} \int_{\hat{t}_{i-1}}^{\hat{t}_i} \big\langle p^i(t), A'(u;t)\delta u\ y(t) - l'(u;t)\delta u\ \big\rangle_{V,V^*} dt + \alpha\langle u, \delta u\rangle, \tag{4.9}$$

where y solves (4.2) and the p^i solve (4.3), (4.4) resp. (4.5), (4.6).

Proof Differentiating $f(u)$ with respect to u in direction δu leads to

$$f'(u)\delta u = \sum_{i=1}^{D} \big\langle C^*\big(Cy(u;\hat{t}_i) - d_i\big), z(\hat{t}_i)\big\rangle_H + \alpha\langle u, \delta u\rangle.$$

For the first summand we get by using (4.4), (4.6) and $z(0)=0$:

$$\begin{aligned}
&\sum_{i=1}^{D}\big\langle C^*\big(Cy(u;\hat{t}_i)-d_i\big), z(\hat{t}_i)\big\rangle_H \\
&\quad=\big\langle -p^D(\hat{t}_D), z(\hat{t}_D)\big\rangle_H+\sum_{i=1}^{D-1}\big\langle -p^i(\hat{t}_i)+p^{i+1}(\hat{t}_i), z(\hat{t}_i)\big\rangle_H+\big\langle p^1(0), z(0)\big\rangle_H \\
&\quad=-\sum_{i=1}^{D}\big(\big\langle p^i(\hat{t}_i), z(\hat{t}_i)\big\rangle_H-\big\langle p^i(\hat{t}_{i-1}), z(\hat{t}_{i-1})\big\rangle_H\big).
\end{aligned} \tag{4.10}$$

Because $p^i \in W([\hat{t}_{i-1},\hat{t}_i], V)$ $(i=1,\dots,D)$ integration by parts can be applied to every summand in (4.10). If further (4.3), (4.5) and (4.8) are used, we get for $i=1,\dots,D$:

$$\begin{aligned}
&\big\langle p^i(\hat{t}_i), z(\hat{t}_i)\big\rangle_H-\big\langle p^i(\hat{t}_{i-1}), z(\hat{t}_{i-1})\big\rangle_H \\
&\quad=\int_{\hat{t}_{i-1}}^{\hat{t}_i}\big(\big\langle z(t), \dot{p}^i(t)\big\rangle_{V,V^*}+\big\langle p^i(t), \dot{z}(t)\big\rangle_{V,V^*}\big)dt \\
&\quad=\int_{\hat{t}_{i-1}}^{\hat{t}_i}\big(\big\langle z(t), A^*(u;t)p^i(t)\big\rangle_{V,V^*}+\big\langle p^i(t), \dot{z}(t)\big\rangle_{V,V^*}\big)dt \\
&\quad=\int_{\hat{t}_{i-1}}^{\hat{t}_i}\big\langle p^i(t), A(u;t)z(t)+\dot{z}(t)\big\rangle_{V,V^*}dt \\
&\quad=-\int_{\hat{t}_{i-1}}^{\hat{t}_i}\big\langle p^i(t), A'(u;t)\delta u\; y(u;t)-l'(u;t)\delta u\big\rangle_{V,V^*}dt,
\end{aligned}$$

which shows the proposition. □

In order to solve the optimization problem numerically, the next step is the discretization. In the first-optimize approach the discretization schemes can be adapted separately for the state and adjoint equations. Section 2.2 showed the difficulties arising in the numerical solution of the state equation due to a non-smooth initial condition. However, this problem gets even more pronounced regarding the adjoint equations. Here, the pointwise observations in the objective function represented by the approximative operator C lead to high-frequency end conditions. Fortunately, the Rannacher smoothing procedure can also be applied for every adjoint p^i, $i=1,\dots,D$ smoothing the non-smooth end conditions

Let $\Delta t=T/n_t$ be the step size of a time discretization and $t_i=i\cdot\Delta t$ $(i=0,\dots,n_t)$ the corresponding grid points. We denote by $Y_k\approx y(t_k)$ and $P_k^i\approx p^i(t_k)$ for $k=1,\dots,m$ (of course the p^i resp. P_k^i only exist on $[\hat{t}_{i-1},\hat{t}_i]$). For simplicity we set $\{\hat{t}_i\}_{i=0}^D\subset\{t_i\}_{i=0}^{n_t}$ and define for every interval $[\hat{t}_{i-1},\hat{t}_i]$ the index set $\mathcal{T}_i:=\{k: t_k\in[\hat{t}_{i-1},\hat{t}_i]\}$ and the indices $i_{min}=\min(\mathcal{T}_i)$ and $i_{max}=\max(\mathcal{T}_i)$.

Definition 4.3 For given weights ω_k^i $(i = 1 \ldots, D,\ k \in \mathcal{T}_i)$ we define the gradient approximation $f'_{FO}(u)\delta u$ by

$$f'_{FO}(u)\delta u = \Delta t \sum_{i=1}^{D} \sum_{k \in \mathcal{T}_i} \omega_k^i \big\langle P_k^i,\, A'(u; t_k)\delta u\ Y_k - l'(u, t_k)\delta u \big\rangle_{V,V*} + \alpha \langle u, \delta u \rangle. \tag{4.11}$$

Remark 4.4 Note that the weights ω_k^i determine the numerical integration rule, e.g., the ‘composite trapezoidal rule’, where $\omega_{i_{min}}^i = \omega_{i_{max}}^i = 0.5$ and $\omega_k^i = 1$ for all other k.

5 First Discretize

We now want to discretize first and use the θ-scheme for the time discretization. For this, we define a time discretization grid $t_0, \ldots, t_m$ with $t_j = j \cdot \Delta t$, $j = 0, \ldots, m$. For simplicity, we assume that the time instances, where market data is available, $\hat{t}_i$ $(i = 1, \ldots, D)$, are a subset of the grid, i.e. there exist subindices such that $\hat{t}_i = t_{k_i}$.

Definition 5.1 For given market data d_i at $\hat{t}_i$ $(i = 1, \ldots, D)$ and $\alpha > 0$, find $y \in (V)^m$ and $u \in \mathcal{U}$, which solve the optimization problem

$$\min_{y \in (V)^m, u \in \mathcal{U}} \tilde{J}(y, u) := \min_{y \in (V)^m, u \in \mathcal{U}} \frac{1}{2} \sum_{i=1}^{D} \left\| C y_{k_i} - d_i \right\|_{\mathcal{H}}^2 + \frac{\alpha}{2} \|u\|^2 \tag{5.1}$$

$$\begin{aligned} s.t. \quad & \frac{1}{\Delta t}(y_{k+1} - y_k) + \theta A(u; t_{k+1}) y_{k+1} + (1 - \theta) A(u; t_k) y_k \\ & \quad - \theta l(u; t_{k+1}) - (1 - \theta) l(u; t_k) = 0, \quad k = 0, \ldots, m - 1, \qquad (5.2) \\ & y_0 = \hat{y}. \end{aligned}$$

A reduced cost function $f_{FD}(u) = \tilde{J}(y(u), u)$ can be defined as in the previous section. Together with the corresponding adjoint equation,

$$\begin{aligned} & \frac{1}{\Delta t}(p_k - p_{k+1}) + \theta A^*(u, t_k) p_k + (1 - \theta) A^*(u, t_k) p_{k+1} \\ & \quad + \sum_{i=1}^{D} C^*(C y_{k_i} - d_i) \mathbb{1}_{k = k_i} = 0, \quad k = m - 1, \ldots, 1, \qquad (5.3) \\ & p_m + \Delta t \theta A^*(u, t_m) p_m = -\Delta t C^*(C y_m - d_D), \end{aligned}$$

which again can be derived via the Lagrangian approach, a gradient representation for the discrete reduced problem can be verified via a discrete sensitivity equation.

Theorem 5.2 *The derivative of $\tilde{f}(u)$ (defined in (3.3)) along a feasible direction δu is given by*

$$f'_{FD}(u)\delta u = \sum_{k=0}^{m-1} \big\langle p_{k+1}, \theta A'(u;t_{k+1})\delta u y_{k+1} + (1-\theta)A'(u;t_k)\delta u y_k \tag{5.4}$$

$$- \theta l'(u;t_{k+1})\delta u - (1-\theta)l'(u;t_k)\delta u \big\rangle_{V,V^*} dt + \alpha\langle u, \delta u\rangle, \tag{5.5}$$

where y solves (5.2) and the p solves (5.3).

6 Comparison and Numerical Results

In this section we discuss the numerical calculation of the gradient of the objective function. Here, the focus is on numerical issues in the time discretization arising in the solution of the adjoint equation.

The numerical results presented below are all based on the Merton model (cf. Example 2.1) and the following setting:

- $\underline{x} = -5, \overline{x} = 5, T_{max} = 2y, r \equiv 3\ \%, \alpha = 0.$
- Market data given at:

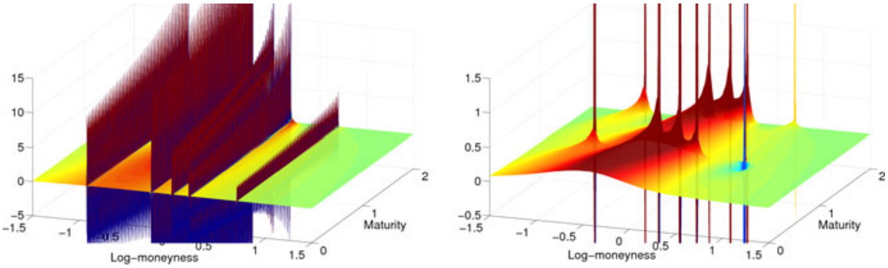
$$(T_i, K_i) \in \big\{\{1,2\} \times \{40\ \%, 80\ \%, 100\ \%, 120\ \%, 200\ \%\}\big\}. \tag{6.1}$$

- Four parameters for Merton's model: $u = (\sigma, \lambda, \mu_J, \sigma_J) \in \mathbb{R}^4$.

The market data call prices are produced with $\tilde{u} = (30\ \%, 50\ \%, 0\ \%, 50\ \%)$ and we choose as a sample parameter $u = (30\ \%, 60\ \%, -80\ \%, 40\ \%)$ to calculate gradients and adjoints. In the abstract setting above, we set $H = L^2(\underline{x}, \overline{x})$, $V = H_0^1(\underline{x}, \overline{x})$ and $\mathcal{U} = \mathbb{R}^4$, whereas, to be precise, the control space is restricted by box constraints.

We already noticed that the pointwise observations in the objective function of the calibration problem (3.1) lead to high-frequency end conditions in the backward adjoint equations.

If we **first optimize**, we are free in the choice of appropriate discretization methods for the state and adjoint equations, separately. As has been shown in Sect. 2.2, a Crank–Nicolson method loses some of its stability properties in the case of non-smooth initial conditions. With regard to the adjoint equation, the problem is far more severe due to appearance of Dirac delta functions spread out over the whole time domain. If a standard Crank–Nicolson method is applied to the adjoint equation (4.3), (4.4) and (4.5), (4.6), this leads to the result illustrated in Fig. 3a ($\Delta x = 0.0025$, $\Delta T = 0.02$). According to the notation of Sect. 4, the adjoint is formally divided into two parts, p^1, p^2, with end conditions at $T = 1$ and $T = 2$, where market data are available. The peaks occurring in these end conditions are not smoothened out, but oscillate strongly over the whole time domain. However,



(A) Adjoint equation: first optimize, then discretize (Crank–Nicolson)

(B) Adjoint equation: first optimize, then discretize (Rannacher time stepping)

Fig. 3 First optimize: solutions of the adjoint equation for ($\Delta T = 0.02$, $\Delta x = 0.0025$)

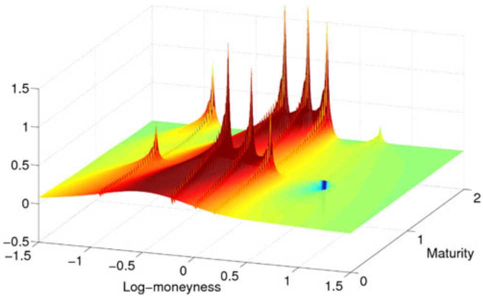


Fig. 4 First discretize: solution of the adjoint equation ($\Delta T = 0.02$, $\Delta x = 0.0025$)

if Rannacher smoothing is applied to the adjoint at each end condition, the corresponding Fig. 3b shows functions p^1, p^2 that are smooth in time.[1]

We now turn to the **first discretize** approach, where we use a Rannacher time stepping scheme for the state equation as proposed in Sect. 5 only in order to smoothen out the slightly nonsmooth initial condition. Figure 4 shows the numerical solution. Note here that the discretization scheme in the state equation automatically yields a Crank–Nicolson method for the adjoint except for the last time step before $T = 0$, where four implicit Euler quartersteps are applied.

Two points are remarkable here. First we note that the peaks at $T = 1$ and $T = 2$ are not as pronounced as in the first-optimize approach. This is due to the fact that the end condition,

$$p_m + \Delta t \theta A^*(u, t_m) p_m = -\Delta t C^*(C y_m - d_D), \tag{6.2}$$

contains a kind of built-in smoothing through the elliptic operator weighted with step size Δt. Hence, the greater the step size, i.e. the more unstable the Crank–Nicolson scheme for non-smooth end conditions, the more pronounced is the

[1] Note the different scaling of Figs. 3a and 3b, causing a cut of the peaks at $T = 1$ and $T = 2$ in Fig. 3b.

Table 3 $L^2(\Omega)$-error at $T = 0$ of the adjoint solution for the three approaches of Figs. 3 and 4 and different time step sizes ΔT and fixed Δx (Reference solution calculated with FO (Rann.) and $\Delta T = 3.125\text{e-}4$, $\Delta x = 0.0025$)

Discretization		$L^2(\Omega)$-error at $T = 0$					
Δx	ΔT	FO (Rann.)	Ratio	FO (C.-N.)	Ratio	FD (Rann.)	Ratio
0.0025	0.04	4.69e-5		1.61e+0		2.55e-3	
	0.02	1.19e-5	3.9	1.05e+0	1.5	1.29e-3	2.0
	0.01	2.99e-6	4.0	3.91e-1	2.7	6.47e-4	2.0
	0.005	7.47e-7	4.0	1.22e-1	3.2	3.24e-4	2.0
	0.0025	1.85e-7	4.0	1.21e-1	1.0	1.62e-4	2.0

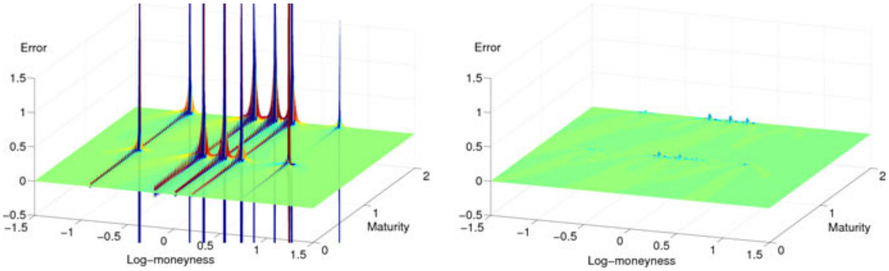
smoothing. However,—and this is the second point—there are still oscillations observable through the whole time domain, which are due to the missing Rannacher smoothing steps after each peak for the adjoint equation.

This is also shown in Table 3, where we compare the numerical solution of the adjoint equation at the last time instance $T = 0$ with a reference solution calculated with first-optimize including Rannacher smoothing on a very fine time grid ($\Delta T = 3.125\text{e-}4$, $\Delta x = 0.0025$). This is done for the three methods shown in Figs. 3 and 4 and for different step sizes ΔT (Δx fixed).

The term 'ratio' again shows the factor of decrease in the $L^2(\Omega)$-error when the number of discretization steps in time is doubled ($\Omega = [-3, 3]$). In the last column of the table the ratio implies only a linear convergence with respect to the step size ΔT. However, the first-optimize approach using Crank–Nicolson shows nearly no improvement of the error for a refined time grid. As expected, the Rannacher smoothing steps in the first optimize approach preserve the quadratic convergence, where the order of magnitude of the error compared to the first discretize approach is significant.

In addition, the error results of Table 3 are visualized in Fig. 5. We omit the result for first-optimize with Crank–Nicolson since this is not competitive. Again we calculate the reference solution for the adjoint on a fine grid ($\Delta T = 3.125\text{e-}4$, $\Delta x = 0.0025$). Figure 5a then shows the error of the adjoint on a coarse time grid ($\Delta T = 0.02$, $\Delta x = 0.0025$), where we first discretized the state equation with Rannacher time stepping for the initial data only and the adjoint equation is then solved by the resulting Crank–Nicolson scheme with four implicit Euler quartersteps in the last time step. The oscillations that are visible in Fig. 4 are now observable more clearly. However, Fig. 5b shows a remarkably smaller error, when we first optimize and then use the Rannacher smoothing for the adjoint.

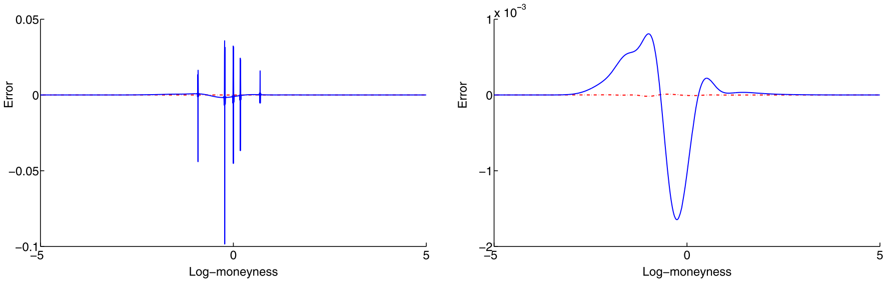
We further want to point out the effect of the four implicit quartersteps, when we discretize first. Figure 6 shows the error at time $T = 0.02$, i.e. the last time step before the implicit steps are applied. Where the first discretize approach (continuous line) shows strong oscillations, the first optimize approach (dotted line) is quite smooth. But the oscillations are then smoothed out in the last time step, observable in Fig. 6b.



(A) Error adjoint: first discretize, then optimize

(B) Error adjoint: first optimize, then discretize (Rannacher time stepping)

Fig. 5 Difference between reference adjoint and the adjoints on coarser grids for first discretize (**a**) and first optimize (**b**), resp.



(A) Error at $T = 0.02$

(B) Error at $T = 0$

Fig. 6 Difference between reference adjoint and the adjoints on coarser grids for first discretize and first optimize (*dotted line*), resp., at time $T = 0$ and at time $T = 0.02$

This observation also motivates a different approach that can be found in [14]. They proposed to change the discretization scheme for the state equation in such way that the resulting scheme for the adjoint equation in the first discretize approach automatically leads to a stabilized version.

Finally, we are interested in the gradient of our problem. Given the parameter vector $u \in \mathbb{R}^4$ for the Merton model, a reference gradient, ∇f_{ref}, is calculated on a fine grid via the first optimize approach with Rannacher smoothing.[2] Table 4 shows the relative errors between this reference gradient and the gradient for the three approaches on several coarser time grids. It is observable that, especially for very coarse time steps ΔT, the first-optimize approach with Rannacher smoothing is by far the best one. However, first-discretize also leads to acceptable results, especially for finer grids. However, it could happen that the oscillations that are observable in the adjoint equation may sum up to zero for certain examples.

[2] Note that the relative difference between reference gradient ($\Delta T = 3.125\text{e-}4$, $\Delta x = 0.0025$) based on first optimize-Rannacher and first discretize-Rannacher is 7.47e-009.

Table 4 Relative gradient errors for the three approaches of Figs. 3 and 4 and different time step sizes ΔT and fixed Δx with control $u \in \mathbb{R}^4$

Discretization		$\|\nabla f_{ref} - \nabla f_{disc}\|/\|\nabla f_{ref}\|$					
Δx	ΔT	FO (Rann.)	Ratio	FO (C.-N.)	Ratio	FD (Rann.)	Ratio
0.0025	0.04	3.50e-5		6.53e-1		6.72e-4	
	0.02	8.48e-6	4.1	1.96e-1	3.3	1.83e-3	0.4
	0.01	2.12e-6	4.0	2.78e-1	0.7	3.23e-4	5.7
	0.005	5.27e-7	4.0	2.95e-1	0.9	1.22e-5	26.5
	0.0025	1.38e-7	3.8	2.94e-1	1.0	5.43e-7	22.5

7 Conclusion and Outlook

In this paper we take a closer look at the gradient computation for a calibration problem in mathematical finance which occurs in finding the optimal parameters of a local volatility model using jump diffusion processes. Basically, two approaches are analyzed: First optimize vs. first discretize. It turns out, that the flexibility of the first optimize approach gives numerically much more stable results since the adjoint equation can be solved by a particular discretization scheme. Numerical results for an example confirm this observation.

References

1. Y. Achdou, O. Pironneau, *Computational Methods for Option Pricing* (SIAM, Philadelphia, 2005)
2. A. Almendral, C. Oosterlee, Numerical valuation of options with jumps in the underlying. Appl. Numer. Math. **53**(1), 1–18 (2005)
3. L. Andersen, J. Andreasen, Jump-diffusion processes: volatility smile fitting and numerical methods for option pricing. Rev. Deriv. Res. **4**(3), 231–262 (2000)
4. O.E. Barndorff-Nielsen, Processes of normal inverse Gaussian type. Finance Stoch. **2**, 41–68 (1997)
5. D. Bates, Jump and stochastic volatility: exchange rate processes implicit in Deutsche Mark options. Rev. Financ. Stud. **9**, 69–107 (1996)
6. F. Black, M. Scholes, The pricing of options and corporate liabilities. J. Polit. Econ. **81**(3), 637–654 (1973)
7. P. Carr, H. Geman, D. Madan, M. Yor, The fine structure of asset returns: an empirical investigation. J. Bus. **75**, 305–332 (2002)
8. N. Chriss, *Black–Scholes and Beyond: Option Pricing Models* (Irwin/McGraw Hill, New York, 1997)
9. R. Cont, P. Tankov, *Financial Modelling with Jump Processes* (Chapman and Hall, London, 2004)
10. R. Dautray, J.-L. Lions, *Evolution Problems I*, Mathematical Analysis and Numerical Methods for Science and Technology, vol. 5 (Springer, Berlin, 1992)
11. E. Derman, I. Kani, The volatility smile and its implied tree. Risk **7**(2), 139–145 (1994)
12. B. Dupire, Pricing with a smile. Risk **7**, 1–10 (1994)
13. M.B. Giles, R. Carter, Convergence analysis of Crank-Nicolson and Rannacher time-marching. J. Comput. Finance **9**(4), 89–112 (2006)

14. C. Goll, R. Rannacher, W. Wollner, The damped Crank-Nicolson time-marching scheme for the adaptive solution of the Black–Scholes equation. J. Comput. Finance (to appear)
15. W. Gong, M. Hinze, Z. Zhou, A priori error analysis for finite element approximation of parabolic optimal control problems with pointwise control. Hamburger Beiträge zur Angewandten Mathematik, Preprint 2011-22, 22 pp., 2011
16. S.L. Heston, A closed-form solution for options with stochastic volatility with applications to bond and currency options. Rev. Financ. Stud. **6**, 327–343 (1993)
17. J.C. Hull, *Options, Futures, and Other Derivatives*, 7th edn. (Prentice Hall, Upper Saddle River, 2008)
18. J.C. Hull, A.D. White, The pricing of options on assets with stochastic volatilities. J. Finance **42**(2), 281–300 (1987)
19. J.P. Morgan, Pricing exotics under smile. Risk **November**, 72–75 (1999)
20. S.G. Kou, A jump-diffusion model for option pricing. Manag. Sci. **48**(8), 1086–1101 (2002)
21. A. Lipton, The vol smile problem. Risk **November**, 61–65 (2002)
22. R.C. Merton, Theory of rational option pricing. Bell J. Econ. Manag. Sci. **4**, 141–183 (1973)
23. R.C. Merton, Option pricing when underlying stock returns are discontinuous. J. Financ. Econ. **3**(1–2), 125–144 (1976)
24. R. Rannacher, Finite element solution of diffusion problems with irregular data. Numer. Math. **43**, 309–327 (1984)
25. E.W. Sachs, M. Schu, Reduced order models in PIDE constrained optimization. Control Cybern. **39**, 661–675 (2010)
26. E.W. Sachs, M. Schu, A priori error estimates for reduced order models in finance. Math. Model. Numer. Anal. **47**, 449–469 (2013)
27. E.W. Sachs, A.K. Strauss, Efficient solution of a partial integro-differential equation in finance. Appl. Numer. Math. **58**, 1687–1703 (2008)
28. W. Schoutens, *Lévy-Processes in Finance* (Wiley, Chichester, 2003)
29. L.R. Scott, Finite element convergence for singular data. Numer. Math. **21**(4), 317–327 (1973)
30. E.M. Stein, J.C. Stein, Stock price distributions with stochastic volatility: an analytic approach. Rev. Financ. Stud. **4**(4), 727–752 (1991)
31. P. Wilmott, J. Dewynne, J. Howison, *Option Pricing: Mathematical Models and Computation* (Oxford Financial Press, Oxford, 1993)

Numerical Analysis of POD A-posteriori Error Estimation for Optimal Control

Alina Studinger and Stefan Volkwein

Abstract In this paper a linear-quadratic optimal control problem governed by a parabolic equation is considered. To solve this problem numerically a reduced-order approach based on proper orthogonal decomposition (POD) is applied. The error between the POD suboptimal control and the optimal control of the original problem is controlled by an a-posteriori error analysis. In this paper the authors focus on testing the a-posteriori estimate's validity by means of numerical examples. An intensive study of the consequences of certain choices that can be made within the POD basis determination process is carried out and the findings are discussed.

Keywords Optimal control · Model reduction · Proper orthogonal decomposition · A-posteriori error estimates · Primal-dual active set strategy

Mathematics Subject Classification (2010) Primary 35K90 · 49K20 · Secondary 65K05

1 Introduction

Optimal control problems for partial differential equation are often hard to tackle numerically because their discretization leads to large scale optimization problems. Therefore, different techniques of model reduction were developed to approximate these problems by smaller ones that are tractable with less effort. Among them, proper orthogonal decomposition (POD) [20] and balanced truncation [3] seem to

The second author gratefully acknowledges support by the German Science Fund DFG grant VO 1658/2-1 *A-posteriori-POD Error Estimators for Nonlinear Optimal Control Problems governed by Partial Differential Equations*.

S. Volkwein (✉) · A. Studinger
Department of Mathematics and Statistics, University of Constance, Universitätsstraße 10, D-78457 Konstanz, Germany
e-mail: Stefan.Volkwein@uni-konstanz.de

A. Studinger
e-mail: astudinger@web.de

K. Bredies et al. (eds.), *Control and Optimization with PDE Constraints*,
International Series of Numerical Mathematics 164,
DOI 10.1007/978-3-0348-0631-2_8, © Springer Basel 2013

be most widely used in the context of optimal control. Recently, optimal control problems are also treated by the reduced basis method; we refer, e.g., to [5, 6, 17].

POD is based on projecting the dynamical system onto subspaces of basis elements that express characteristics of the expected solution. This is in contrast to, e.g., finite element techniques, where the elements are not correlated to the physical properties of the system they approximate.

In our present work, POD is applied to linear-quadratic optimal control problems. Linear-quadratic problems are interesting in several respects; in particular, since they occur in each level of sequential quadratic programming (SQP) methods; see, e.g., [19] from a general viewpoint and [12, 20] in the context of multilevel reduced-order approximations. We continue the research on POD a-posteriori error analysis; see [12, 13, 20, 23, 25]. Based on a perturbation argument it is derived how far the suboptimal control, computed on the basis of the POD model, is from the (unknown) exact one. Increasing the number of POD ansatz functions leads to more accurate POD suboptimal controls. This idea turns out to be numerically very efficient. It is also successfully applied for other reduced-order approximations; see [22, 26].

Here, we focus on testing the a-posteriori estimate's validity by means of numerical examples. We intensively study the consequences of certain choices that can be made within the POD basis determination process. Let us summarize the key findings here:

- The estimation is very satisfactory and valuable in terms of accuracy, reliability and efficiency.
- Both the primal-dual active set strategy for solving the control constrained optimal control problem and the a-posteriori error estimation for tracking the error work very well for control box constraints. In case of active constraints, we discover numerical convergence of the active sets which is perfect in case of an "optimal" POD basis (computed from the optimal FE solution) and satisfactory for arbitrary bases.
- In order to obtain good POD suboptimal controls it is not sufficient to solely increase the number of used POD basis functions. Increasing the basis rank needs to be combined with basis update strategies, as for example discussed in [1, 2, 15, 25].
- Enriching the snapshot ensemble by snapshots from the adjoint state is essential to obtain good approximations for the control.

This paper is organized as follows: In Sect. 2 we introduce the abstract linear-quadratic optimal control problem and review first-order necessary optimality conditions. The POD method, its application to the optimal control problem and the a-posteriori error estimate are explained in Sect. 3. In Sect. 4 numerical examples are presented and discussed.

2 The Optimal Control Problem

In this section, we introduce a class of linear-quadratic parabolic optimal control problems and recall the associated first-order optimality conditions.

2.1 Problem Formulation

Let V and H be real, separable Hilbert spaces and suppose that V is dense in H with compact embedding. By $\langle\cdot,\cdot\rangle_H$ we denote the inner product in H. The inner product in V is given by a symmetric bounded, coercive, bilinear form $a : V \times V \to \mathbb{R}$:

$$\langle \varphi, \psi\rangle_V = a(\varphi, \psi) \quad \text{for all } \varphi, \psi \in V \tag{2.1}$$

with associated norm $\|\cdot\|_V = \sqrt{a(\cdot\,,\cdot)}$. By identifying H with its dual H' it follows that $V \hookrightarrow H = H' \hookrightarrow V'$, each embedding being continuous and dense. Recall that for $T > 0$ the space $W(0,T)$

$$W(0,T) = \left\{\varphi \in L^2(0,T;V) : \varphi_t \in L^2\big(0,T;V'\big)\right\}$$

is a Hilbert space endowed with the common inner product [4]. When the time t is fixed, the expression $\varphi(t)$ stands for the function $\varphi(t,\cdot)$ considered as a function in Ω only. Let $\mathcal{D}$ be an open and bounded subset in $\mathbb{R}^m$ with $m \in \mathbb{N}$. By U_{ad} we denote the closed, convex and bounded subset

$$U_{\mathsf{ad}} = \left\{u \in L^2(\mathcal{D}) \mid u_a(s) \le u(s) \le u_b(s) \text{ for almost all (f.a.a.) } s \in \mathcal{D}\right\},$$

where $u_a, u_b \in L^2(\mathcal{D})$ satisfy $u_a \le u_b$ almost everywhere (a.e.) in $\mathcal{D}$. For $y_0 \in V$, $f \in L^2(0,T;H)$ and $u \in U_{\mathsf{ad}}$ we consider the linear evolution problem

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\big\langle y(t),\varphi\big\rangle_H + a\big(y(t),\varphi\big) &= \big\langle (f+\mathcal{B}u)(t),\varphi\big\rangle_H \quad && \text{f.a.a. } t \in [0,T],\ \forall \varphi \in V,\\ \big\langle y(0),\varphi\big\rangle_H &= \langle y_0,\varphi\rangle_H && \forall \varphi \in V, \end{aligned} \tag{2.2}$$

where $\mathcal{B} : L^2(\mathcal{D}) \to L^2(0,T;H)$ is a continuous, linear operator.

Example 2.1 Let us present an example for (2.2) which will be studied in our numerical experiments. Suppose that $\Omega \subset \mathbb{R}^2$, is an open and bounded domain with Lipschitz-continuous boundary $\Gamma = \partial\Omega$. For $T > 0$ we set $Q = (0,T) \times \Omega$ and $\Sigma = (0,T) \times \Gamma$. Let $H = L^2(\Omega)$, $V = H^1(\Omega)$ and $\mathcal{D} = \Sigma$. Then, for given control $u \in L^2(\Sigma)$ and initial condition $y_0 \in V$ we consider

$$c_p y_t(t,\boldsymbol{x}) - \Delta y(t,\boldsymbol{x}) = \tilde{f}(t,\boldsymbol{x}) \quad \text{f.a.a. } (t,\boldsymbol{x}) \in Q, \tag{2.3a}$$

$$\frac{\partial y}{\partial n}(t,\boldsymbol{x}) + q y(t,\boldsymbol{x}) = u(t,\boldsymbol{x}) \quad \text{f.a.a. } (t,\boldsymbol{x}) \in \Sigma, \tag{2.3b}$$

$$y(0,\boldsymbol{x}) = y_0(\boldsymbol{x}) \quad \text{f.a.a. } \boldsymbol{x} \in \Omega. \tag{2.3c}$$

In (2.3a) we suppose $c_p > 0$, $q \ge 0$ and $\tilde{f} \in L^2(0,T;H)$. Setting $f = \tilde{f}/c_p$, introducing the bounded bilinear form $a : V \times V \to \mathbb{R}$ by

$$a(\varphi,\psi) = \frac{1}{c_p}\int_\Omega \nabla\varphi(\boldsymbol{x}) \cdot \nabla\psi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \frac{q}{c_p}\int_\Gamma \varphi(\boldsymbol{x})\psi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \quad \text{for } \varphi,\psi \in V$$

and the linear, bounded operator $\mathcal{B} : L^2(\Sigma) \to L^2(0,T;H)$ by

$$\big\langle (\mathcal{B}u)(t), \varphi \big\rangle_H = \frac{1}{c_p} \int_\Gamma u(t,\boldsymbol{x})\varphi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \quad \text{for } \phi \in V,\ t \in (0,T) \text{ a.e.}$$

then the weak formulation of (2.3a)–(2.3c) can be expressed in the form (2.2).

It is known [4] that for every $f \in L^2(0,T;H)$, $u \in L^2(\mathcal{D})$ and $y_0 \in V$ there is a unique weak solution $y \in W(0,T) \cap C([0,T];V)$ satisfying (2.2).

Remark 2.2 Let $\hat{y}_0 \in W(0,T)$ be the unique solution to

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\big\langle \hat{y}_0(t), \varphi \big\rangle_H + a\big(\hat{y}_0(t), \varphi\big) &= \big\langle f(t), \varphi \big\rangle_H \quad && \text{f.a.a. } t \in [0,T],\ \forall \varphi \in V, \\ \big\langle \hat{y}_0(0), \varphi \big\rangle_H &= \langle y_0, \varphi \rangle_H && \forall \varphi \in V. \end{aligned}$$

Moreover, we introduce the linear and bounded operator $\mathcal{S} : L^2(\mathcal{D}) \to W(0,T)$ as follows: $\tilde{y} = \mathcal{S}u \in W(0,T)$ is the unique solution to

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\big\langle \tilde{y}(t), \varphi \big\rangle_H + a\big(\tilde{y}(t), \varphi\big) &= \big\langle (\mathcal{B}u)(t), \varphi \big\rangle_H \quad && \text{f.a.a. } t \in [0,T],\ \forall \varphi \in V, \\ \big\langle \tilde{y}(0), \varphi \big\rangle_H &= 0 && \forall \varphi \in V. \end{aligned}$$

Then, $y = \hat{y}_0 + \mathcal{S}u$ is the weak solution to (2.2).

Next we introduce the cost functional $J : W(0,T) \times L^2(\mathcal{D}) \to \mathbb{R}$ by

$$J(y,u) = \frac{1}{2}\big\| y(T) - y_d \big\|_H^2 + \frac{\gamma}{2}\|u\|^2_{L^2(\mathcal{D})}, \tag{2.4}$$

where $y_d \in H$ holds. Furthermore, $\gamma > 0$ is a regularization parameter.

Remark 2.3 We continue Example 2.1. Then, (2.4) yields the cost functional

$$J(y,u) = \frac{1}{2}\int_\Omega \big| y(T) - y_d \big|^2\,\mathrm{d}\boldsymbol{x} + \frac{\gamma}{2}\int_0^T \int_\Gamma \big| u(t,\boldsymbol{x}) \big|^2\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t$$

for $(y,u) \in W(0,T) \times L^2(\Sigma)$.

The optimal control problem is given by

$$\min J(y,u) \quad \text{subject to (s.t.)} \quad (y,u) \in W(0,T) \times U_{\mathsf{ad}} \text{ solves (2.2).} \tag{$\mathbf{P}$}$$

Applying standard arguments [16] one can prove that there exists a unique optimal solution $\bar{x} = (\bar{y}, \bar{u})$ to (**P**). Throughout this paper, a bar indicates optimality.

2.2 *First-Order Optimality Conditions*

First-order necessary optimality conditions for our parabolic optimal control problem are well known. We briefly recall them here. Suppose that $\bar{x} = (\bar{y}, \bar{u})$ is the optimal solution to (**P**). Then there exists a unique Lagrange-multiplier $\bar{p} \in W(0,T)$ satisfying together with $\bar{x}$ the *first-order necessary optimality conditions*, which consist of the *state equations* (2.2), the *adjoint equations*

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\big\langle \bar{p}(t), \varphi\big\rangle_H + a\big(\bar{p}(t), \varphi\big) &= 0 \quad \text{f.a.a. } t \in [0,T],\ \forall \varphi \in V,\\ \big\langle \bar{p}(T), \varphi\big\rangle_H = \big\langle y_d - \bar{y}(T), \varphi\big\rangle_H & \qquad \forall \varphi \in V \end{aligned} \tag{2.5}$$

and of the *variational inequality*

$$\big\langle \gamma\bar{u} - \mathcal{B}^\star \bar{p}, u - \bar{u}\big\rangle_{L^2(\mathcal{D})} \geq 0 \quad \forall u \in U_{\mathsf{ad}}. \tag{2.6}$$

Here, the linear and bounded operator $\mathcal{B}^\star : L^2(0,T;H) \to L^2(\mathcal{D})' \sim L^2(\mathcal{D})$ stands for the dual operator of $\mathcal{B}$ satisfying

$$\langle \mathcal{B}u, \varphi\rangle_{L^2(0,T;H)} = \big\langle u, \mathcal{B}^\star\varphi\big\rangle_{L^2(\mathcal{D})} = \big\langle \mathcal{B}^\star\varphi, u\big\rangle_{L^2(\mathcal{D})}$$

for all $(u,\varphi) \in L^2(\mathcal{D}) \times L^2(0,T;H)$.

Remark 2.4 We continue the discussion of Example 2.1 and Remark 2.3. The adjoint equations (2.5) are given by

$$\begin{aligned} -c_p\bar{p}_t(t,\boldsymbol{x}) - \Delta\bar{p}(t,\boldsymbol{x}) &= 0 && \text{f.a.a. } (t,\boldsymbol{x}) \in Q,\\ \frac{\partial\bar{p}}{\partial n}(t,\boldsymbol{x}) + q\bar{p}(t,\boldsymbol{x}) &= 0 && \text{f.a.a. } (t,\boldsymbol{x}) \in \Sigma,\\ \bar{p}(T,\boldsymbol{x}) &= y_d(\boldsymbol{x}) - \bar{y}(T,\boldsymbol{x}) && \text{f.a.a. } \boldsymbol{x} \in \Omega. \end{aligned}$$

Moreover, the variational inequality (2.6) has the form

$$\int_0^T\int_\Gamma \big(\gamma\bar{u}(t,\boldsymbol{x}) - \bar{p}(t,\boldsymbol{x})\big)\big(u(t,\boldsymbol{x}) - \bar{u}(t,\boldsymbol{x})\big)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \geq 0 \quad \text{for all } u \in U_{\mathsf{ad}}$$

and $\mathcal{B}^\star\bar{p}$ is given by $(\mathcal{B}^\star\bar{p})(t) = (\tau_\Gamma\bar{p})(t)$ f.a.a. $t \in [0,T]$, where $\tau_\Gamma : V \to L^2(\Gamma)$ denotes the common trace operator.

3 The POD Galerkin Discretization

Problem (**P**) is an infinite-dimensional problem. Therefore, we have to discretize (**P**) for its numerical solution. For the discretization of the spatial variable we apply a POD Galerkin approximation, which is discussed now. Let X denote either the space H or the space V.

3.1 The POD Method

Let an arbitrary $u \in L^2(\mathcal{D})$ be chosen such that the corresponding state variable $y = \hat{y}_0 + \mathcal{S}u \in W(0,T)$ belongs to $C([0,T];V) \hookrightarrow C([0,T];X)$. Then,

$$\mathcal{V} = \operatorname{span}\big\{y(t) \mid t \in [0,T]\big\} \subseteq V \subset X. \tag{3.1}$$

If $y_0 \neq 0$ holds, then $\operatorname{span}\{y_0\} \subset \mathcal{V}$ and $d = \dim \mathcal{V} \in [1, \infty]$, but $\mathcal{V}$ may have infinite dimension. We define a bounded linear operator $\mathcal{Y} : L^2(0,T) \to X$ by

$$\mathcal{Y}\varphi = \int_0^T \varphi(t)y(t)\,\mathrm{d}t \quad \text{for } \varphi \in L^2(0,T).$$

Its Hilbert space adjoint $\mathcal{Y}^\star : X \to L^2(0,T)$ satisfying

$$\langle \mathcal{Y}\varphi, z\rangle_X = \big\langle \varphi, \mathcal{Y}^\star z\big\rangle_{L^2(0,T)} \quad \text{for } (\varphi, z) \in L^2(0,T) \times X$$

is given by $(\mathcal{Y}^\star z)(t) = \langle z, y(t)\rangle_X$ for $z \in X$ and f.a.a. $t \in [0,T]$. The bounded linear operator $\mathcal{R} = \mathcal{Y}\mathcal{Y}^\star : X \to \mathcal{V} \subset X$ has the form

$$\mathcal{R}z = \int_0^T \big\langle z, y(t)\big\rangle_X \, y(t)\,\mathrm{d}t \quad \text{for } z \in X. \tag{3.2}$$

Moreover, let $\mathcal{K} = \mathcal{Y}^\star\mathcal{Y} : L^2(0,T) \to L^2(0,T)$ be defined by

$$(\mathcal{K}\varphi)(t) = \int_0^T \big\langle y(\tau), y(t)\big\rangle_X \, \varphi(\tau)\,\mathrm{d}\tau \quad \text{for } \varphi \in L^2(0,T).$$

It is known [11, Sect. 3] that the operator $\mathcal{R}$ is self-adjoint, compact and non-negative. Thus, that there exists a complete orthonormal basis $\{\psi_i\}_{i=1}^d$ for $\mathcal{V} = \operatorname{range}(\mathcal{R}) \subseteq V$ and a sequence $\{\lambda_i\}_{i=1}^d$ of real numbers such that

$$\mathcal{R}\psi_i = \lambda_i\psi_i \quad \text{for } i = 1, \ldots, d \quad \text{and} \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0. \tag{3.3}$$

Remark 3.1

(1) The linear, bounded, compact and self-adjoint operator $\mathcal{K}$ has the same eigenvalues $\{\lambda_i\}_{i=1}^d$ as the operator $\mathcal{R}$. For all $\lambda_i > 0$ the corresponding eigenfunctions of $\mathcal{K}$ are given by

$$v_i(t) = \frac{1}{\sqrt{\lambda_i}}\big(\mathcal{Y}^*\psi_i\big)(t) = \frac{1}{\sqrt{\lambda_i}}\big\langle \psi_i, y(t)\big\rangle_X \quad \text{f.a.a. } t \in [0,T] \text{ and } 1 \leq i \leq \ell.$$

(2) Notice that $\mathcal{V} \subset V$ implies $\psi_i \in V$ for $1 \leq i \leq \ell$.

For $\ell \leq d$ the eigenvalues and eigenfunctions of $\mathcal{R}$ solve

$$\min \int_0^T \left\| y(t) - \sum_{i=1}^{\ell} \big\langle y(t), \psi_i \big\rangle_X \psi_i \right\|_X^2 \mathrm{d}t \quad \text{s.t. } \langle \psi_j, \psi_i \rangle_X = \delta_{ij}, 1 \leq i, j \leq \ell. \tag{3.4}$$

In particular,

$$\int_0^T \left\| y(t) - \sum_{i=1}^{\ell} \big\langle y(t), \psi_i \big\rangle_X \psi_i \right\|_X^2 \mathrm{d}t = \sum_{i=\ell+1}^{d} \lambda_i .$$

3.2 The Discrete POD Method

In real computations, we do not have the whole trajectory $y(t)$ for all $t \in [0, T]$. For that purpose let $0 \leq t_1 < t_2 < \cdots < t_n \leq T$ be a given time grid and let $y_j^h \approx y(t_j)$ denote approximations in a finite-dimensional space $X^h \subset X$ for y at time instance t_j, $j = 1, \ldots, n$. We set $\mathcal{V}^n = \operatorname{span}\{y_1^h, \ldots, y_n^h\}$ with $d^n = \dim \mathcal{V}^n \leq n$. Then, for given $\ell \leq n$ we consider the problem

$$\min \sum_{j=1}^{n} \alpha_j \left\| y_j^h - \sum_{i=1}^{\ell} \big\langle y_j^h, \psi_i^n \big\rangle_X \psi_i^n \right\|_X^2 \quad \text{s.t. } \big\langle \psi_i^n, \psi_j^n \big\rangle_X = \delta_{ij},\ 1 \leq i, j \leq \ell \tag{3.5}$$

instead of (3.4). In (3.5), the α_j's stand for the trapezoidal weights

$$\alpha_1 = \frac{t_2 - t_1}{2}, \qquad \alpha_j = \frac{t_{j+1} - t_{j-1}}{2} \quad \text{for } 2 \leq j \leq n-1, \qquad \alpha_n = \frac{t_n - t_{n-1}}{2}.$$

The solution to (3.5) is given by the solution to the eigenvalue problem

$$\mathcal{R}^n \psi_i^n = \sum_{j=1}^{n} \alpha_j \big\langle y_j^h, \psi_i^n \big\rangle_X y_j^h = \lambda_i^n \psi_i^n, \quad i = 1, \ldots, \ell,$$

where $\mathcal{R}^n : X^h \to \mathcal{V}^n \subset V$ is a linear, bounded, compact, self-adjoint and non-negative operator. Thus, there are an orthonormal set $\{\psi_i^n\}_{i=1}^n$ of eigenfunctions and corresponding non-negative eigenvalues $\{\lambda_i^n\}_{i=1}^n$ satisfying

$$\mathcal{R}^n \psi_i^n = \lambda_i^n \psi_i^n, \qquad \lambda_1^n \geq \lambda_2^n \geq \cdots \geq \lambda_{d^n}^n > \lambda_{d^n+1}^n = \cdots = \lambda_n^n = 0. \tag{3.6}$$

We refer to [14] for the relationship between (3.3) and (3.6).

Remark 3.2 Let X^h be given by the subset $\operatorname{span}\{\varphi_1, \ldots, \varphi_m\} \subset X$, where the φ_i's are assumed to be linearly independent in X. Then we have

$$y_j^h(\boldsymbol{x}) = \sum_{i=1}^{m} Y_{ij} \varphi_i(\boldsymbol{x}) \in X^h \quad \text{for } \boldsymbol{x} \in \Omega \text{ and } j = 1, \ldots, n$$

with real coefficients Y_{ij}. In this case the POD basis functions are given by

$$\psi_j^n(\boldsymbol{x}) = \sum_{i=1}^m \Psi_{ij}\varphi_i(\boldsymbol{x}) \in X^h \quad \text{for } \boldsymbol{x} \in \Omega \text{ and } j = 1, \ldots, \ell$$

with real coefficients Ψ_{ij}. Then we have to determine the coefficient matrix $\Psi = ((\Psi_{ij})) \in \mathbb{R}^{m\times\ell}$. For that purpose we define $Y = ((Y_{ij})) \in \mathbb{R}^{m\times n}$ and $W = ((\langle\varphi_j, \varphi_i\rangle_X)) \in \mathbb{R}^{m\times m}$. Moreover, we define the diagonal matrix $D = \operatorname{diag}(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^{n\times n}$ and set $\bar{Y} = W^{1/2}YD^{1/2} \in \mathbb{R}^{m\times n}$. Then $\Psi = [u_1, \ldots, u_\ell]$ can be computed as follows (see, e.g., [24, Sect. 1.3])

(1) Solve the $m \times m$ eigenvalue problem

$$\bar{Y}\bar{Y}^T \mathfrak{u}_i = \lambda_i \mathfrak{u}_i, \quad 1 \le i \le \ell, \quad \text{with } \mathfrak{u}_i^T \mathfrak{u}_j = \delta_{ij},\ 1 \le i, j \le \ell,$$

for the largest eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_\ell > 0$ and compute $u_i = W^{-1/2}\mathfrak{u}_i$. Since $\bar{Y}\bar{Y}^T = W^{1/2}YDY^TW^{1/2}$ holds, this variant is often numerically expensive, especially for $m \gg n$.

(2) Solve the $n \times n$ eigenvalue problem

$$\bar{Y}^T\bar{Y}\mathfrak{v}_i = \lambda_i \mathfrak{v}_i, \quad 1 \le i \le \ell, \quad \text{with } \mathfrak{v}_i^T \mathfrak{v}_j = \delta_{ij},\ 1 \le i, j \le \ell, \tag{3.7}$$

for the largest eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_\ell > 0$ and set $u_i = YD^{1/2}\mathfrak{v}_i/\sqrt{\lambda_i}$. To solve (3.7) we apply the MATLAB routine `eigs` and call this variant 'eigs' in Sect. 4. Note that $\bar{Y}^T\bar{Y} = D^{1/2}Y^TWYD^{1/2}$ holds and D is a diagonal matrix. Since we do not have to compute $W^{1/2}$, this variant is very attractive for $n \le m$. We will apply this approach in our numerical experiments.

(3) Compute the singular value decomposition (SVD) of $\bar{Y}$, i.e., determine orthonormal vectors $\{\mathfrak{u}_i\}_{i=1}^\ell$ in $\mathbb{R}^m$ and $\{\mathfrak{v}_i\}_{i=1}^\ell$ in $\mathbb{R}^n$ associated with the largest singular values $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_\ell > 0$ satisfying

$$\bar{Y}\mathfrak{v}_i = \sigma_i \mathfrak{u}_i, \quad \bar{Y}^T \mathfrak{u}_i = \sigma_i \mathfrak{v}_i, \quad 1 \le i \le \ell$$

(see, e.g., [18]). It follows that $\lambda_i = \sigma_i^2$ and $u_i = W^{-1/2}\mathfrak{u}_i$. Since this variant is based on the SVD, we call this variant 'SVD' in Sect. 4. Although the computation of $W^{1/2}$ is costly, the SVD is known to be more stable. This is due to the fact that the products of $\bar{Y}$ and $\bar{Y}^T$ squares the condition number of the problem compared to the SVD.

3.3 POD Galerkin Approximation for (P)

Let $y = \hat{y}_0 + \mathcal{S}u$ be the state associated with some control $u \in U$, and let $\mathcal{V}$ be given as in (3.1). We fix ℓ with $\ell \le d$ and compute the first ℓ POD basis functions $\psi_1, \ldots, \psi_\ell \in V$ by solving either $\mathcal{R}\psi_i = \lambda_i\psi_i$ or $\mathcal{K}v_i = \lambda v_i$ for $i = 1, \ldots, \ell$ (see

Remark 3.1). Then we define

$$V^\ell = \operatorname{span}\{\psi_1, \ldots, \psi_\ell\} \subset V.$$

Endowed with the topology in V it follows that V^ℓ is a Hilbert space. The POD Galerkin scheme for the state equation (2.2) leads to the following linear problem: determine a function $y^\ell(t) \in V^\ell$ such that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\big\langle y^\ell(t), \psi\big\rangle_H + a\big(y^\ell(t), \psi\big) &= \big\langle (f + \mathcal{B}u)(t), \psi\big\rangle_H \\ &\qquad \text{f.a.a. } t \in [0, T],\ \forall \psi \in V^\ell, \\ \big\langle y^\ell(0), \psi\big\rangle_H &= \langle y_0, \psi\rangle_H \quad \forall \psi \in V^\ell. \end{aligned} \tag{3.8}$$

For every $f \in L^2(0, T; H)$, $u \in L^2(\mathcal{D})$, $y_0 \in V$ and for every $\ell \in \mathbb{N}$ problem (3.8) admits a unique solution $y^\ell \in H^1(0, T; V^\ell)$; see [10, Proposition 3.4]. From $V^\ell \hookrightarrow V$ it follows that $y^\ell \in W(0, T)$ holds.

The POD Galerkin approximation for (**P**) is given by problem

$$\min J\big(y^\ell, u\big) \quad \text{s.t.} \quad \big(y^\ell, u\big) \in H^1\big(0, T; V^\ell\big) \times U_{\mathsf{ad}} \text{ solves (3.8).} \tag{$\mathbf{P}^\ell$}$$

Problem ($\mathbf{P}^\ell$) admits a unique optimal solution $\bar{x}^\ell = (\bar{y}^\ell, \bar{u}^\ell)$ that is interpreted as a suboptimal solution to (**P**). First-order necessary optimality conditions for ($\mathbf{P}^\ell$) are given by the *state equation* (3.8) with $u = \bar{u}^\ell$, the *adjoint equation*

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\big\langle \bar{p}^\ell(t), \psi\big\rangle_H + a\big(\bar{p}^\ell(t), \psi\big) &= 0 \quad \text{f.a.a. } t \in [0, T],\ \forall \psi \in V^\ell, \\ \big\langle \bar{p}^\ell(T), \psi\big\rangle_H &= \big\langle y_d - \bar{y}^\ell(T), \psi\big\rangle_H \quad \forall \psi \in V^\ell. \end{aligned} \tag{3.9}$$

and *the variational inequality*

$$\big\langle \gamma \bar{u}^\ell - \mathcal{B}^\star \bar{p}^\ell, u - \bar{u}^\ell\big\rangle_{L^2(\mathcal{D})} \geq 0 \quad \text{for all } u \in U_{\mathsf{ad}}.$$

To solve ($\mathbf{P}^\ell$) we apply a primal-dual active set strategy, which converges locally superlinearly [7]. Its mesh-independence is proved in [8, 9].

3.4 A-posteriori Error Estimate for the POD Approximation

In this subsection we present the a-posteriori error estimate for the control variable. The result is taken from [23, Theorem 4.11].

Theorem 3.3 *Suppose that* $(\bar{y}, \bar{u})$ *is the solution to* (**P**)*. For an arbitrary* $\ell \leq d$ *let* $(\bar{y}^\ell, \bar{u}^\ell)$ *be the optimal solution to* ($\mathbf{P}^\ell$)*. Let* $\tilde{y} = \hat{y}_0 + \mathcal{S}\bar{u}^\ell$ *and* $\tilde{p} = \tilde{p}(\bar{u}^\ell)$ *be the solution to the associated adjoint equation*

$$
\begin{aligned}
-\frac{\mathrm{d}}{\mathrm{d}t}\big\langle \tilde p(t),\varphi\big\rangle_H + a\big(\tilde p(t),\varphi\big) &= 0, && t\in[0,T],\ \forall\varphi\in V,\\
\big\langle \tilde p(T),\varphi\big\rangle_H &= \big\langle y_d-\tilde y(T),\psi\big\rangle_H && \forall\varphi\in V.
\end{aligned}
\tag{3.10}
$$

Define the residual function $\zeta^\ell\in L^2(\mathcal{D})$ by

$$
\zeta^\ell(s)=\begin{cases}
[(\gamma\bar u^\ell-\mathcal{B}^\star\tilde p)(s)]_- & \textit{on } \mathcal{A}_-^\ell=\{s\in\mathcal{D}\mid \bar u^\ell(s)=u_a(s)\},\\
[(\gamma\bar u^\ell-\mathcal{B}^\star\tilde p)(s)]_+ & \textit{on } \mathcal{A}_+^\ell=\{s\in\mathcal{D}\mid \bar u^\ell(s)=u_b(s)\},\\
-(\gamma\bar u^\ell-\mathcal{B}^\star\tilde p)(s) & \textit{on } \mathcal{J}^\ell=\mathcal{D}\setminus(\mathcal{A}_-^\ell\cup\mathcal{A}_+^\ell),
\end{cases}
\tag{3.11}
$$

with $[r]_-=-\min(0,r)$ *and* $[r]_+=\max(0,r)$. *Then*

$$
\big\|\bar u-\bar u^\ell\big\|_{L^2(\mathcal{D})}\le\frac{1}{\gamma}\,\big\|\zeta^\ell\big\|_{L^2(\mathcal{D})}.
$$

Remark 3.4

(1) Notice that $\tilde y$ and $\tilde p$ must be taken as the solutions to the (full) state and adjoint equation, respectively, not of their POD-approximations.

2) In [23] sufficient conditions are presented that $\lim_{\ell\to\infty}\|\zeta^\ell\|_{L^2(\mathcal{D})}=0$. Thus, $\|\zeta^\ell\|_{L^2(\mathcal{D})}$ can be expected smaller than any $\varepsilon>0$ provided that ℓ is taken sufficiently large. Motivated by this result, we set up Algorithm 1.

(3) Notice that the presented error estimate holds for time-variant, linear-quadratic optimal control problems. For recent extension to nonlinear problems we refer to [13] and to [12, 20], where the presented error estimate is utilized in a multilevel SQP algorithm.

(4) To improve the approximation quality of the POD basis, we can combine the a-posteriori analysis with basis update strategies; see [25].

Algorithm 1 (POD reduced-order method with a-posteriori estimator)

```
Choose $u \in U_{\mathsf{ad}}$, an initial number $\ell$ for POD ansatz functions, a maximal number $\ell^{\max} > \ell$ of POD ansatz functions, $\varepsilon > 0$; compute $y = \hat y_0 + \mathcal{S}u$.
Determine a POD basis of rank $\ell^{\max}$ utilizing the state $y = \hat y_0 + \mathcal{S}u$.
repeat
    Build the reduced-order problem ($\mathbf{P}^\ell$) of rank $\ell \le \ell^{\max}$.
    Compute the suboptimal control $\bar u^\ell$.
    Determine $\tilde y = \hat y_0 + \mathcal{S}\bar u^\ell$, $\tilde p$ (see (3.10)) as well as $\zeta^\ell$ (see (3.11)).
    if $\|\zeta^\ell\|_{L^2(\mathcal{D})} < \varepsilon$ or $\ell = \ell^{\max}$ then
        Return $\ell$, suboptimal control $\bar u^\ell$ and STOP.
    else
        Set $\ell = \ell + 1$.
    end if
until $\ell > \ell^{\max}$
```

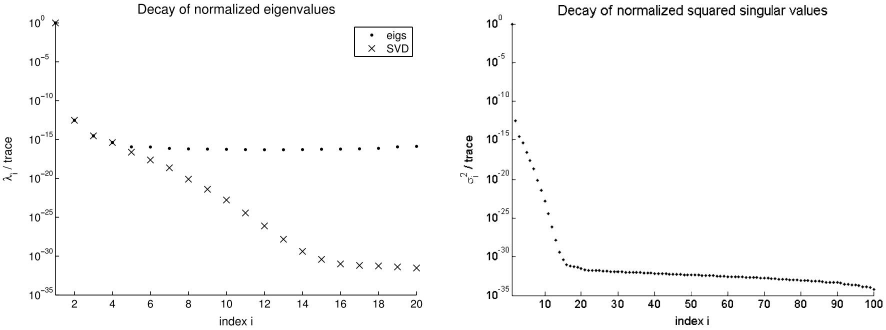


Fig. 1 Run 4.1: Decay of $\bar{\lambda}_i$ for 'eigs' (*left plot*) and $\bar{\sigma}_i^2$ for 'SVD' (*right plot*) with $X = H$

4 Numerical Experiments

This chapter is devoted to numerical test examples. First, we turn to the numerical solution of a given parabolic PDE. We pursue the two different Galerkin approaches, namely the finite element (FE) Galerkin method and the POD Galerkin technique, and compare the results in order to see some different implementation choices that can be taken, to get a sense of how good these approximate solutions are and to point out some advantages and drawbacks of the considered methods. From solving one linear parabolic equation we move on to applying the POD-Galerkin ansatz to (**P**). Hereby, we especially focus on testing the accuracy and efficiency of the POD a-posteriori estimator reviewed in Sect. 3.4. For the implementation we use the MATLAB software package (R2010a). In all examples we choose the spatial domain $\Omega = (0,1) \times (0,1) \subset \mathbb{R}^2$. The time interval of consideration will always be $[0,T] = [0,1]$. The time integration is carried out by the implicit Euler method with an equidistant time grid $0 = t_0 < t_1 < \cdots < t_n = T$, where $t_i = i\Delta t$ and $\Delta t = 1/n$.

Run 4.1 (Heat equation) In (2.3a)–(2.3c) we choose the data $y_0 \equiv 0$, $c_p = 1$, $q = 0$ and $\tilde{f}(t,\boldsymbol{x}) = \cos(2\pi x_1)\cos(2\pi x_2)(1 + 8\pi^2 t)$, $(t,\boldsymbol{x}) \in Q$ and $\boldsymbol{x} = (x_1, x_2)$. Then, the exact solution is $y_{ex}(t,\boldsymbol{x}) = t\cos(2\pi x_1)\cos(2\pi x_2)$ for $(t,\boldsymbol{x}) \in Q$. For the FE method we choose piecewise quadratic elements resulting in $m = 665$ spatial degrees of freedom. The time increment was chosen to be $\Delta t = 0.01$ and we have $n = 101$. Notice that the discretization error with respect to the spatial and the time variable is of the same size $\mathcal{O}(\Delta t)$. To compute the POD basis we compare the variants 'eigs' and 'SVD'; see Remark 3.2. Choosing $X = H$ in Sect. 3.1 the rapid decays of the normalized eigenvalues

$$\bar{\lambda}_i = \frac{\lambda_i}{\sum_{j=1}^n \lambda_j} = \frac{\lambda_i}{\operatorname{trace}(\bar{Y}^T\bar{Y})}$$

and normalized squared singular values $\bar{\sigma}_i^2 = \bar{\lambda}_i$ are presented in Fig. 1. We observe that in the beginning the eigenvalues are equal in deed, whereas the eigenvalues computed with the SVD keep decreasing when the eigenvalues for 'eigs' stagnate

Table 1 Run 4.1: Absolute & relative errors and λ_ℓ for different ℓ and for 'eigs' and 'SVD' using $X = H$

ℓ	Variant 'eigs'			Variant 'SVD'		
	$\mathcal{E}^y_{\mathsf{abs}}(\ell)$	$\mathcal{E}^y_{\mathsf{rel}}(\ell)$	λ_ℓ	$\mathcal{E}^y_{\mathsf{abs}}(\ell)$	$\mathcal{E}^y_{\mathsf{rel}}(\ell)$	λ_ℓ
1	$1.3 \cdot 10^{-07}$	$2.1 \cdot 10^{-06}$	$8.5 \cdot 10^{-00}$	$1.3 \cdot 10^{-07}$	$2.1 \cdot 10^{-06}$	$8.5 \cdot 10^{-00}$
2	$1.4 \cdot 10^{-08}$	$5.8 \cdot 10^{-07}$	$2.5 \cdot 10^{-12}$	$1.4 \cdot 10^{-08}$	$5.8 \cdot 10^{-07}$	$2.5 \cdot 10^{-12}$
5	$4.3 \cdot 10^{-10}$	$1.1 \cdot 10^{-08}$	$1.1 \cdot 10^{-15}$	$4.3 \cdot 10^{-10}$	$1.1 \cdot 10^{-08}$	$2.0 \cdot 10^{-16}$
10	$3.2 \cdot 10^{-13}$	$5.4 \cdot 10^{-12}$	$4.6 \cdot 10^{-16}$	$1.6 \cdot 10^{-13}$	$2.7 \cdot 10^{-12}$	$1.4 \cdot 10^{-22}$
11	$6.0 \cdot 10^{-13}$	$7.8 \cdot 10^{-12}$	$4.4 \cdot 10^{-16}$	$2.2 \cdot 10^{-14}$	$3.4 \cdot 10^{-13}$	$3.1 \cdot 10^{-24}$
14	$7.7 \cdot 10^{-12}$	$3.1 \cdot 10^{-11}$	$4.4 \cdot 10^{-16}$	$5.2 \cdot 10^{-16}$	$2.5 \cdot 10^{-15}$	$3.5 \cdot 10^{-29}$
15	$9.9 \cdot 10^{-12}$	$4.0 \cdot 10^{-11}$	$4.6 \cdot 10^{-16}$	$5.5 \cdot 10^{-16}$	$2.5 \cdot 10^{-15}$	$3.2 \cdot 10^{-30}$

at the order of machine precision. The difference between the 'eigs' and the 'SVD' variant shows already for small ℓ in this example due to the extremely rapid decay of eigenvalues, see Table 1, where

$$\mathcal{E}^y_{\mathsf{abs}}(\ell) = \sum_{j=1}^{n} \alpha_j \left\| y^h(t_j) - y^\ell(t_j) \right\|_H, \qquad \mathcal{E}^y_{\mathsf{rel}}(\ell) = \sum_{j=1}^{n} \alpha_j \, \frac{\| y^h(t_j) - y^\ell(t_j) \|_H}{\| y^h(t_j) \|_H}$$

stands for absolute and relative error between the FE and POD solution. In fact, the solution space of the PDE is one-dimensional, since the exact solution at time t is given by a multiple of $\cos(2\pi x_1)\cos(2\pi x_2)$ by the factor t. This behavior is already captured by one mode/basis function. Due to the inaccuracy of the numerical method for determining the snapshots, the snapshot matrix Y representing the solution space has a rank greater than one. Thus the first POD basis function ψ_1 is not an exact multiple of $\cos(2\pi x_1)\cos(2\pi x_2)$, and the dynamics of the PDE can not be described comprehensively with only one POD basis function. Hence, increasing the rank of the POD basis leads the approximation quality to rise.

The fact that the eigenvalues for 'eigs' increase starting from $\ell = 14$ instead of continuing to decrease like for 'SVD' illustrate that the SVD is more stable than the eigenvalue solver. Since the matrix $\bar{Y}^T \bar{Y}$ is symmetric, positive semi-definite, it should only have real non-negative eigenvalues. Due to rounding errors, eigenvalues that are nearly of the size of the machine precision ($\texttt{eps} = 2.2204 \cdot 10^{-16}$) can be mistakenly determined as negative or complex by `eigs.m`. For SVD the eigenvalues keep on decreasing since they are computed as squares of the obtained singular values. This leads also to a monotone decrease of the quantities $\mathcal{E}^y_{\mathsf{abs}}(\ell)$ and $\mathcal{E}^y_{\mathsf{rel}}(\ell)$ for the SVD, whereas 'eigs' yields a stagnation of the error for $\ell \geq 14$. In case of 'eigs' we observe that the stagnation of the quantities $\mathcal{E}^y_{\mathsf{abs}}(\ell)$ and $\mathcal{E}^y_{\mathsf{rel}}(\ell)$ happens before the corresponding eigenvalues stagnate. This is due to the loss of the W-orthonormality of the POD basis vectors. Analytically, we have

$$\Psi^T_{\cdot,1:\ell} W \Psi_{\cdot,1:\ell} - I_\ell \overset{!}{=} 0,$$

Table 2 Run 4.1: Spectral norm $\|\Psi^T_{\cdot,1:\ell} W \Psi_{\cdot,1:\ell} - I_\ell\|_2$ for different ℓ and for 'eigs' & 'SVD' using $X = H$

ℓ	$\|\Psi^T_{\cdot,1:\ell} W \Psi_{\cdot,1:\ell} - I_\ell\|_2$	
	Variant 'eigs'	Variant 'SVD'
1	$8.12 \cdot 10^{-16}$	$2.66 \cdot 10^{-15}$
2	$3.51 \cdot 10^{-05}$	$3.98 \cdot 10^{-15}$
5	$9.95 \cdot 10^{-01}$	$3.94 \cdot 10^{-15}$
10	$9.99 \cdot 10^{-01}$	$1.05 \cdot 10^{-14}$

where I_ℓ denotes the ℓ by ℓ identity matrix and $\Psi_{\cdot,1:\ell}$ contains the first ℓ columns of ψ. We estimate the spectral norm $\|\Psi^T_{\cdot,1:\ell} W \Psi_{\cdot,1:\ell} - I_\ell\|_2$ by utilizing the MATLAB routine `normest`; compare Table 2. We observe that the 'SVD' approach fulfills the W-orthogonality far better than the 'eigs' variant. From Tables 1 and 2, we can deduce that especially for higher POD basis rank the SVD is more stable and accurate. However, we should mention that the 'SVD' variant is more costly than the 'eigs' variant, especially if the number of spatial degrees of freedom m is much bigger that the number n of time steps.

Run 4.2 (Unconstrained optimal control) In the context of Example 2.1, Remark 2.3 and Remark 2.4 we choose $\gamma = 10^{-2}$, $c_p = 10$, $q = 0.01$ and $\tilde{f} \equiv 0$. The initial condition is $y_0(\boldsymbol{x}) = 3 - 4(x_2 - 0.5)^2$ and the desired state is $y_d(\boldsymbol{x}) = 2 + 2\,|2x_1 - x_2|$ for $\boldsymbol{x} = (x_1, x_2) \in \Omega$. Choosing $u_a = -\infty = -u_b$ we have $U_{\mathsf{ad}} = L^2(\Sigma)$. We make use of the MATLAB PDE toolbox for the spatial discretization with piecewise linear, continuous finite elements (P_1-Elements) with maximal edge length $h_{\max} = 0.06$ and thus $N_{FE} = 498$ degrees of freedom. For the implicit Euler method we choose the step size $\Delta t = 0.004$. The FE optimal control $\bar{u}^h$ is presented at all times in Fig. 2. The different tested ROM runs vary in the way the POD basis is determined:

(1) To generate the snapshots for the POD method we have to solve (2.3a)–(2.3c) for a reference control $u = u_{\mathsf{ref}}$. We consider two different reference controls:
   (1a) $u^1_{\mathsf{ref}}(t, \boldsymbol{x}) = \exp(t)(\frac{1}{2}\,|2x_1 - x_2| + \frac{1}{3}\,(\sin(\pi x_2) - 1))$ for $(t, \boldsymbol{x}) \in \Sigma$, $\boldsymbol{x} = (x_1, x_2)$ (see Fig. 3);
   (1b) $u^2_{\mathsf{ref}}(t, \boldsymbol{x}) = \bar{u}^h(t, \boldsymbol{x})$ for $(t, \boldsymbol{x}) \in \Sigma$ (see Fig. 2).
   The reference control u^1_{ref} is plotted in Fig. 3. Notice that u^1_{ref} shares some behavior of the optimal solution u^2_{ref}. In the thesis [21] also the reference control $u^0_{\mathsf{ref}} = 0$ is chosen, which leads—compared to u^1_{ref} to slightly worse numerical results.

(2) The snapshot ensemble to be represented well by the POD basis can now be taken from the solution y_{ref} of the state equation with $u = u^i_{\mathsf{ref}}$, $i = 1, 2$, which is called Variant 1. If we want/need to enrich the approximation space, we also solve the adjoint equation with $y = y_{\mathsf{ref}}$ and then consider a snapshot ensemble consisting of snapshots from both the state and the adjoint equation. This approach is called Variant 2. Let us note that another possibility would be to use two different bases which is not considered in this paper.

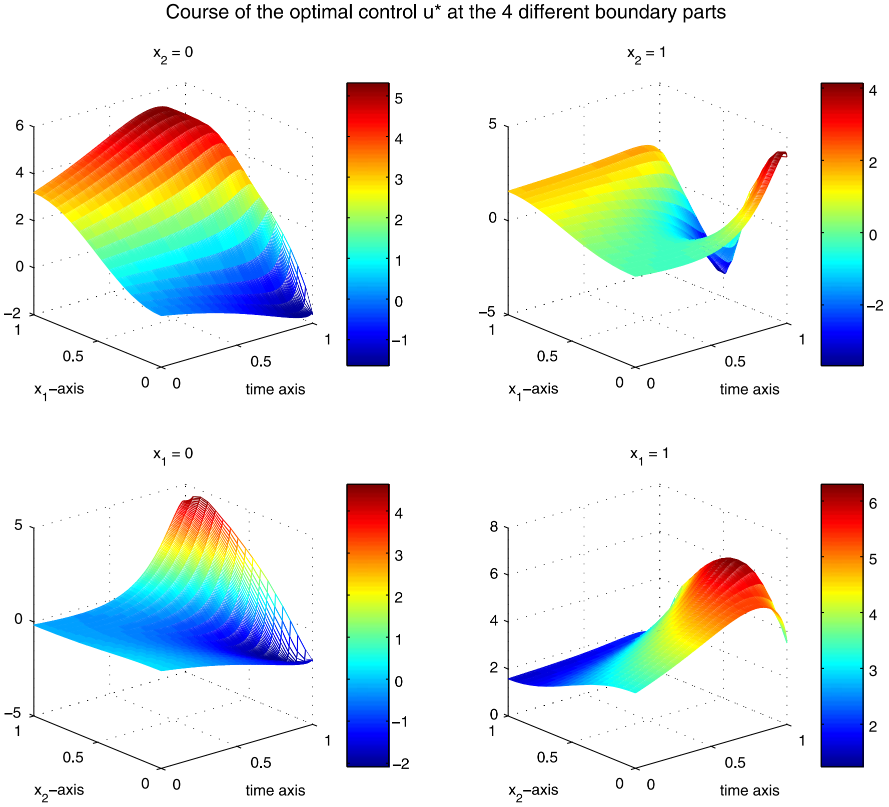


Fig. 2 Run 4.2: FE optimal control $\bar{u}^h(t, \boldsymbol{x})$ for $\boldsymbol{x} = (x_1, 0) \in \Gamma$ (*upper left plot*), $\boldsymbol{x} = (x_1, 1) \in \Gamma$ (*upper right plot*), $\boldsymbol{x} = (0, x_2) \in \Gamma$ (*lower left plot*), $\boldsymbol{x} = (1, x_2) \in \Gamma$ (*lower right plot*)

(3) For the POD basis computation we choose
  (3a) 'eigs' includes solving $\bar{Y}^T \bar{Y} \mathfrak{v} = \lambda \mathfrak{v}$ with $\bar{Y} = W^{1/2} Y D^{1/2}$;
  (3b) 'SVD' involves the singular value decomposition of $\bar{Y}$.

Moreover, the POD basis can be computed for the choices $X = H$ or $X = V$. First, we choose u^1_{ref} for the snapshot generation and use 'eigs' to determine the POD basis based on a snapshot ensemble from both the state and the adjoint equation (Variant 2). In order to get a first idea of how many POD basis functions we should use in the POD-Galerkin ansatz for the state and the adjoint state variable to compute snapshots for the state and the associated adjoint equation using the reference control u^1_{ref}. All snapshots are utilized to compute one POD basis. Then, we look at the decay of the eigenvalues; see Fig. 4. Naturally, in the case of the V-norm the eigenvalues decay slower than with the discrete H-norm. The decay plot shows where the eigenvalues stagnate. Usually, from that number of POD basis functions on, we can not further or significantly improve the approximation errors any more. Theoretically, increasing the POD basis rank leads to a decrease in approximation error

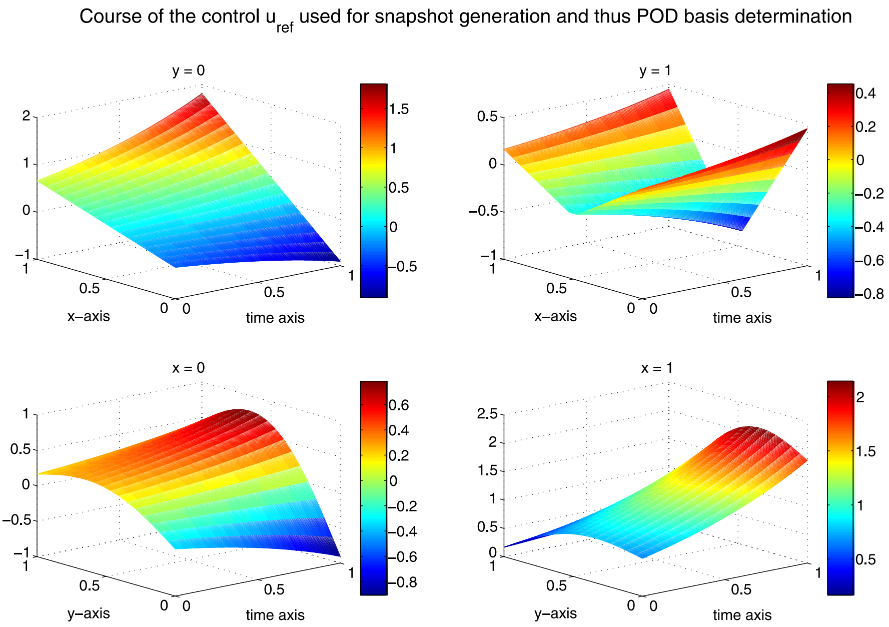


Fig. 3 Run 4.2: u^1_{ref} for $\boldsymbol{x} = (x_1, 0) \in \Gamma$ (*upper left plot*), $\boldsymbol{x} = (x_1, 1) \in \Gamma$ (*upper right plot*), $\boldsymbol{x} = (0, x_2) \in \Gamma$ (*lower left plot*), $\boldsymbol{x} = (1, x_2) \in \Gamma$ (*lower right plot*)

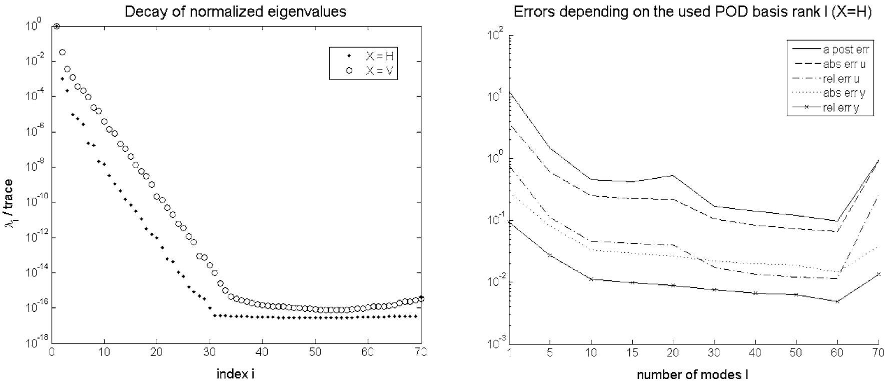


Fig. 4 Run 4.2: Decay of $\bar{\lambda}_i$ for $X = H$ and $X = V$ (*left plot*) and decay of the a-posteriori error estimator, the absolute as well as the relative errors using u^1_{ref} and 'eigs'

values. Nevertheless, we have to pay attention if we use the not so stable method 'eigs' for POD basis computation. The instability can be detected in the right-hand side plot of Fig. 4 for $\ell > 60$. Note that due to the slower decay of eigenvalues with the H^1-norm implementations those errors are still decreasing for up to $\ell = 70$ POD basis functions. The instability sets in later.

Table 3 Run 4.2: A-posteriori estimator, absolute and relative errors in the control variable for $X = H$ and $X = V$ and for different ℓ using u^1_{ref} an 'eigs'

ℓ	$X = H$			$X = V$		
	$\lVert\zeta^\ell\rVert/\gamma$	$\mathcal{E}^u_{\text{abs}}(\ell)$	$\mathcal{E}^u_{\text{rel}}(\ell)$	$\lVert\zeta^\ell\rVert/\gamma$	$\mathcal{E}^u_{\text{abs}}(\ell)$	$\mathcal{E}^u_{\text{rel}}(\ell)$
1	$1.2 \cdot 10^{+1}$	$3.8 \cdot 10^{-0}$	$7.8 \cdot 10^{-1}$	$1.2 \cdot 10^{+1}$	$3.8 \cdot 10^{-0}$	$7.8 \cdot 10^{-1}$
5	$1.5 \cdot 10^{-0}$	$6.1 \cdot 10^{-1}$	$1.1 \cdot 10^{-1}$	$2.4 \cdot 10^{-0}$	$9.8 \cdot 10^{-1}$	$1.8 \cdot 10^{-1}$
10	$4.5 \cdot 10^{-1}$	$2.5 \cdot 10^{-1}$	$4.5 \cdot 10^{-2}$	$4.4 \cdot 10^{-1}$	$2.5 \cdot 10^{-1}$	$4.6 \cdot 10^{-2}$
20	$5.3 \cdot 10^{-1}$	$2.2 \cdot 10^{-1}$	$4.1 \cdot 10^{-2}$	$5.5 \cdot 10^{-1}$	$2.1 \cdot 10^{-1}$	$3.9 \cdot 10^{-2}$
30	$1.7 \cdot 10^{-1}$	$1.1 \cdot 10^{-1}$	$1.8 \cdot 10^{-2}$	$1.7 \cdot 10^{-1}$	$1.1 \cdot 10^{-1}$	$1.8 \cdot 10^{-2}$
50	$1.2 \cdot 10^{-1}$	$7.4 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$	$1.2 \cdot 10^{-1}$	$7.5 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$
60	$9.9 \cdot 10^{-2}$	$6.6 \cdot 10^{-2}$	$1.1 \cdot 10^{-2}$	$5.8 \cdot 10^{-2}$	$4.2 \cdot 10^{-2}$	$6.8 \cdot 10^{-3}$
70	$9.4 \cdot 10^{-1}$	$9.3 \cdot 10^{-1}$	$2.6 \cdot 10^{-1}$	$5.2 \cdot 10^{-2}$	$3.9 \cdot 10^{-2}$	$6.2 \cdot 10^{-3}$

Let us define the quantities

$$\mathcal{E}^u_{\text{abs}}(\ell) = \sum_{j=0}^{n} \alpha_j \left\lVert \bar{u}^h(t_j) - \bar{u}^\ell(t_j) \right\rVert_{L^2(\Gamma)},$$

$$\mathcal{E}^u_{\text{rel}}(\ell) = \sum_{j=0}^{n} \alpha_j \frac{\lVert \bar{u}^h(t_j) - \bar{u}^\ell(t_j) \rVert_{L^2(\Gamma)}}{\lVert \bar{u}^h(t_j) \rVert_{L^2(\Gamma)}}.$$

The errors which occurred between the (sub-)optimal controls within the first two ROM runs compared to the FE based approaces are listed in Table 3. The obtained results for the POD suboptimal controls are not satisfying. The problem is that increasing the number of utilized POD basis functions does not yield better results, it even leads to meaningless results due to the instability of the 'eigs' method. Even if we consider the 'SVD' approach, the error values do not continue to decrease significantly. They somehow stagnate which is of course better than with the 'eigs' method, but still does not yield a satisfying approximation quality. This is due to the fact that the POD basis is chosen poorly. The generated snapshots for basis determination do not reflect the dynamics of the optimally controlled trajectory, since the reference control is not chosen well enough. Next, we select a better—somehow "optimal"—admissible reference control u_{ref} for snapshot generation, namely the FE optimal control. The POD basis is now determined with $u^2_{\text{ref}} = \bar{u}^h$, $X = V$ and snapshots from both the state and the adjoint equation. Using 'SVD' Table 4 presents the deviation of the POD suboptimal controls/states from the FE optimal solutions depending on the number ℓ of used POD basis functions. From Table 4 we can conclude that the POD basis should not be chosen arbitrarily if we want to obtain a very good approximation quality. The POD based solver with somehow "optimal" reference control u^2_{ref} (Table 4) yields considerably better results than with more or less arbitrary reference control u^1_{ref} (Table 3). Table 4 emphasizes the good quality of the POD a-posteriori error estimator. We can also observe that the a posteriori es-

Table 4 Run 4.2: A-posteriori estimator, absolute and relative errors in the control and state variable for $X = V$ and for different ℓ using u^2_{ref}

ℓ	$\|\zeta^\ell\|/\gamma$	$\mathcal{E}^u_{\mathsf{abs}}(\ell)$	$\mathcal{E}^u_{\mathsf{rel}}(\ell)$	$\mathcal{E}^y_{\mathsf{abs}}(\ell)$	$\mathcal{E}^y_{\mathsf{rel}}(\ell)$
1	$1.1 \cdot 10^{+01}$	$3.4 \cdot 10^{-00}$	$6.9 \cdot 10^{-01}$	$1.7 \cdot 10^{-01}$	$5.8 \cdot 10^{-02}$
5	$2.4 \cdot 10^{-00}$	$1.1 \cdot 10^{-00}$	$2.2 \cdot 10^{-01}$	$5.2 \cdot 10^{-02}$	$1.8 \cdot 10^{-02}$
15	$2.5 \cdot 10^{-02}$	$2.5 \cdot 10^{-02}$	$5.4 \cdot 10^{-03}$	$3.6 \cdot 10^{-04}$	$1.2 \cdot 10^{-04}$
20	$2.8 \cdot 10^{-03}$	$2.8 \cdot 10^{-03}$	$6.0 \cdot 10^{-04}$	$5.1 \cdot 10^{-05}$	$1.8 \cdot 10^{-05}$
40	$2.1 \cdot 10^{-06}$	$2.1 \cdot 10^{-06}$	$4.5 \cdot 10^{-07}$	$1.9 \cdot 10^{-08}$	$6.4 \cdot 10^{-09}$
60	$8.0 \cdot 10^{-10}$	$8.0 \cdot 10^{-10}$	$1.8 \cdot 10^{-10}$	$5.4 \cdot 10^{-12}$	$1.9 \cdot 10^{-12}$
70	$7.0 \cdot 10^{-12}$	$7.0 \cdot 10^{-12}$	$1.5 \cdot 10^{-12}$	$5.0 \cdot 10^{-14}$	$1.7 \cdot 10^{-14}$
90	$3.8 \cdot 10^{-13}$	$1.9 \cdot 10^{-13}$	$3.7 \cdot 10^{-14}$	$2.9 \cdot 10^{-14}$	$9.8 \cdot 10^{-15}$

Table 5 Run 4.2: A-posteriori estimator, absolute error in the control variable and normalized eigenvalues for $X = V$, for the variants 'SVD' as well as 'eigs' and for different ℓ using u^2_{ref}

ℓ	'SVD'			'eigs'		
	$\|\zeta^\ell\|/\gamma$	$\mathcal{E}^u_{\mathsf{abs}}(\ell)$	$\frac{\lambda_\ell}{\mathrm{tr}(\bar{Y}\bar{Y}^T)}$	$\|\zeta^\ell\|/\gamma$	$\mathcal{E}^u_{\mathsf{abs}}(\ell)$	$\frac{\lambda_\ell}{\mathrm{tr}(\bar{Y}\bar{Y}^T)}$
40	$2.1 \cdot 10^{-06}$	$2.1 \cdot 10^{-06}$	$1.1 \cdot 10^{-14}$	$2.1 \cdot 10^{-06}$	$2.1 \cdot 10^{-06}$	$1.1 \cdot 10^{-14}$
60	$8.0 \cdot 10^{-10}$	$8.0 \cdot 10^{-10}$	$5.1 \cdot 10^{-22}$	$2.5 \cdot 10^{-08}$	$2.5 \cdot 10^{-08}$	$1.7 \cdot 10^{-16}$
70	$7.0 \cdot 10^{-12}$	$7.0 \cdot 10^{-12}$	$8.1 \cdot 10^{-26}$	$2.0 \cdot 10^{-09}$	$2.0 \cdot 10^{-09}$	$8.6 \cdot 10^{-17}$
90	$3.8 \cdot 10^{-13}$	$1.9 \cdot 10^{-13}$	$3.1 \cdot 10^{-30}$	$2.7 \cdot 10^{-09}$	$2.5 \cdot 10^{-09}$	$1.1 \cdot 10^{-16}$

timator constitutes a reliable upper bound if the 'eigs' approach was taken for basis determination, see Table 5. As long as there is still enough new information content, meaning that the eigenvalues are still "big" enough so that rounding errors do not jeopardize the decreasing order of eigenvalues and the nearly W-orthogonality, than the 'eigs' and 'SVD' approaches yield the same eigenvalues and the same suboptimal solutions. This is the case up to ℓ scarcely above 40. From then on, the POD basis determination using 'eigs' is not stable any more and does not yield approximation errors as good as the 'SVD' approach, see Table 5. Table 6 gives a summary of the CPU times. Notice that $\ell = 20$ POD basis functions are sufficient for an approximation of the exact optimal control by the POD suboptimal control, since the POD solutions cannot be significantly better than the (piecewise linear) FE solution since it is based on FE snapshots and FE matrices. The FE discretization error cannot be overcome by the POD solutions. In case of $\ell = 20$ and 'eigs', the overall CPU time needed to compute the POD suboptimal control is 5.15 seconds and thus about *240 times smaller* than the CPU time needed to compute the "truth"/FE optimal control.

Finally, we compare the approximation quality of the POD basis if only snapshots from the state (Variant 1) or if snapshots from the state *and* the adjoint equation (Variant 2) are utilized. Even with "optimal" reference control $u^2_{\mathsf{ref}} = \bar{u}^h$, snapshots based solely on the state equation are not sufficient for getting good results with

Table 6 Run 4.2: CPU times in seconds

	CPU time in s
FE solver	1240.87
ROM solver	
snapshot generation (state & adjoint)	1.62
POD basis computation with…	
…eigs ($\ell = 90$)	1.75
…SVD ($\ell = 90$)	6.57
PDASS incl. assembly of all matrices with…	
…$\ell = 20$	1.78
…$\ell = 90$	29.41
a-posteriori error estimation	2.81

the POD ansatz; see Table 7. From Table 7 we conclude that the inclusion of adjoint information into the snapshot ensemble according to Variant 2 is essential to obtain good approximations for the controls. This is due to the fact that the optimality condition directly relates the control onto the adjoint state variable. Hence, it is important to also capture the dynamics of the adjoint equation in order to have a good snapshot ensemble and thus a good POD basis. This coincides with theoretical results, see [10]. The eigenvalues decay slower for $X = V$ than for $X = H$, since there is more information from the snapshot ensemble that gets incorporated into the POD basis. That is why the error values decay slower for small ℓ. Nevertheless, for higher ℓ this higher information content leads to more "stability" and thus monotonously decreasing error values instead of severe oscillations. Summarizing, the choice of $X = V$ and the snapshot ensemble from both the state and the adjoint equation leads to the best performance of the POD-Galerkin ansatz for solving the optimal control problem.

Table 7 Run 4.2: Absolute errors in the control variable for different choices of the snapshot ensemble, for X, for 'eigs' and for different ℓ

ℓ	$\\|\bar{u}^h - \bar{u}^\ell\\|_{L^2(\Sigma)}$			
	Variant 1, $X = V$	Variant 2, $X = V$	Variant 1, $X = H$	Variant 2, $X = H$
1	$3.4 \cdot 10^{-0}$	$3.4 \cdot 10^{-0}$	$3.4 \cdot 10^{-0}$	$3.4 \cdot 10^{-0}$
5	$1.1 \cdot 10^{-0}$	$1.1 \cdot 10^{-0}$	$1.0 \cdot 10^{-0}$	$9.8 \cdot 10^{-1}$
10	$8.8 \cdot 10^{-1}$	$1.7 \cdot 10^{-1}$	$8.5 \cdot 10^{-1}$	$5.9 \cdot 10^{-2}$
20	$6.5 \cdot 10^{-1}$	$2.8 \cdot 10^{-3}$	$6.5 \cdot 10^{-1}$	$1.7 \cdot 10^{-3}$
30	$6.0 \cdot 10^{-1}$	$1.2 \cdot 10^{-4}$	$6.0 \cdot 10^{-1}$	$9.9 \cdot 10^{-5}$
50	$5.1 \cdot 10^{-1}$	$1.1 \cdot 10^{-7}$	$5.1 \cdot 10^{-1}$	$2.4 \cdot 10^{-7}$
60	$5.0 \cdot 10^{-1}$	$2.5 \cdot 10^{-8}$	$1.7 \cdot 10^{+2}$	$1.8 \cdot 10^{-8}$
70	$5.0 \cdot 10^{-1}$	$2.0 \cdot 10^{-9}$	$5.3 \cdot 10^{-1}$	$2.0 \cdot 10^{-7}$

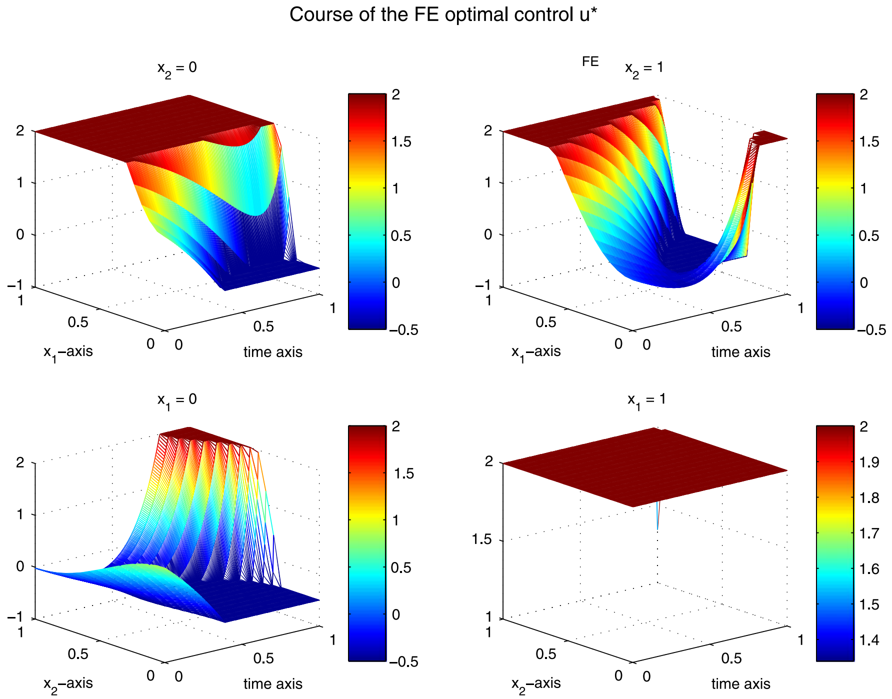


Fig. 5 Run 4.3: FE optimal control $\bar{u}^h(t, \boldsymbol{x})$ for $\boldsymbol{x} = (x_1, 0) \in \Gamma$ (*upper left plot*), $\boldsymbol{x} = (x_1, 1) \in \Gamma$ (*upper right plot*), $\boldsymbol{x} = (0, x_2) \in \Gamma$ (*lower left plot*), $\boldsymbol{x} = (1, x_2) \in \Gamma$ (*lower right plot*)

Run 4.3 (Constrained optimal control) We take the same configuration as in Run 4.2, but now we choose the control constraints $u_a \equiv -0.5$ and $u_b \equiv 2$. Like in Run 4.2 we make use of the MATLAB PDE toolbox for the spatial discretization with piecewise linear, continuous finite elements with 498 degrees of freedom. For temporal discretization, we use the equidistant time increment $\Delta t = 0.004$. The FE solver needs 5 iterations of the primal-dual active set strategy and requires 3435.82 seconds. The optimal control, $\bar{u}^h$ is displayed in Fig. 5. We now test if the implemented ROM solver works properly when there are active box constraints given for the control u. For this, we discuss the results from two different ROM runs. They only vary in the choice of the reference control, first we take the reference control $u^1_{\mathsf{ref}}(t, \boldsymbol{x}) = \exp(t)(\frac{1}{2}|2x_1 - x_2| + \frac{\sin(\pi x_2)}{3} - \frac{1}{3})$ and second we take the FE optimal control $\bar{u}^h$ for snapshot generation. In both cases we use $X = V$, the snapshot ensemble from both state and adjoint equation and the 'eigs' method for POD basis computation.

Note that $u^1_{\mathsf{ref}} \notin U_{\mathsf{ad}}$ holds. Nevertheless, we will see that we still get good results, for higher ℓ the errors are even smaller than in the unrestricted case with this reference control. In Table 8 we present the POD a-posteriori error estimate, the absolute and the relative error in the control variable for different number ℓ of POD basis functions. Taking more POD basis function, into the POD-Galerkin ansatz of

Table 8 Run 4.3: A-posteriori estimator, absolute and relative errors for the control and state variable for $X = V$, different number ℓ of POD basis functions and for the reference control u^1_{ref}

ℓ	$\lVert\zeta^\ell\rVert/\gamma$	$\mathcal{E}^u_{\mathsf{abs}}(\ell)$	$\mathcal{E}^u_{\mathsf{rel}}(\ell)$	$\mathcal{E}^y_{\mathsf{abs}}(\ell)$	$\mathcal{E}^y_{\mathsf{rel}}(\ell)$
1	$1.2 \cdot 10^{+1}$	$2.2 \cdot 10^{-0}$	$7.6 \cdot 10^{-1}$	$1.9 \cdot 10^{-1}$	$6.7 \cdot 10^{-2}$
5	$1.3 \cdot 10^{-0}$	$9.1 \cdot 10^{-1}$	$3.0 \cdot 10^{-1}$	$9.6 \cdot 10^{-2}$	$3.3 \cdot 10^{-2}$
10	$6.2 \cdot 10^{-1}$	$3.7 \cdot 10^{-1}$	$1.3 \cdot 10^{-1}$	$6.7 \cdot 10^{-2}$	$2.3 \cdot 10^{-2}$
20	$5.9 \cdot 10^{-1}$	$3.2 \cdot 10^{-1}$	$1.1 \cdot 10^{-1}$	$3.8 \cdot 10^{-2}$	$1.3 \cdot 10^{-2}$
30	$1.2 \cdot 10^{-1}$	$7.7 \cdot 10^{-2}$	$2.6 \cdot 10^{-2}$	$2.1 \cdot 10^{-2}$	$7.1 \cdot 10^{-3}$
50	$1.9 \cdot 10^{-2}$	$1.7 \cdot 10^{-2}$	$5.5 \cdot 10^{-3}$	$1.5 \cdot 10^{-2}$	$5.2 \cdot 10^{-3}$
60	$1.4 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$	$3.8 \cdot 10^{-3}$	$1.3 \cdot 10^{-2}$	$4.3 \cdot 10^{-3}$
70	$1.2 \cdot 10^{-2}$	$1.1 \cdot 10^{-2}$	$3.5 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}$	$4.3 \cdot 10^{-3}$
90	$1.1 \cdot 10^{-2}$	$9.7 \cdot 10^{-3}$	$3.2 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}$	$4.2 \cdot 10^{-3}$

Table 9 Run 4.3: Number of differences in restricted node values of the POD suboptimal controls $\bar{u}^\ell$ in comparison to the FE optimal control $\bar{u}^h$ using u^1_{ref} and $X = V$

ℓ	5	10	15	20	30	40	50	60	70	80	90	100
diffua	998	818	548	596	106	27	24	19	18	18	17	15
diffub	700	457	412	415	112	31	33	25	23	23	24	24

the state and adjoint state variable makes the errors for the control variable descend. This behavior also becomes evident for the absolute deviation of the POD suboptimal state y^ℓ from the FE optimal state $\bar{y}^h$ as well as for the relative error values.

The FE optimal control is restricted by u_a at a total amount of 3110 boundary nodes over all times and prescribed by u_b at a total amount of 8410 boundary nodes over all times. In Table 9 we can see how many (diffua) boundary node values are determined by u_a in either the POD suboptimal control or in the FE optimal control, but not in both. The differently restricted boundary node values by u_b are counted and stated in diffub. This table shows that the POD suboptimal control $\bar{u}^\ell$ and the FE optimal control $\bar{u}^h$ get restricted/fixed equally at an increasing number of boundary nodes with increasing number ℓ of POD basis functions used to compute $\bar{u}^\ell$. Note that e.g. for $\ell = 100$ there are only 15 differences in lower restricted node values meaning that more than 99 % of the 3110 FE restricted values (by u_a) are replicated by the ROM optimal control solver. The same ratio holds true for the values which are fixed by u_b.

Now we utilize snapshots generated by the reference control $u_{\mathsf{ref}} = \bar{u}^h$. The very good approximation quality can be seen from Table 10. Again, we check if the values of the POD suboptimal control are restricted at the same boundary nodes as the FE optimal control and how this changes depending on the number ℓ of used POD basis functions, see Table 11. The active sets within the reduced order method coincide with those within the FE approach for $\ell \geq 50$. Compared to the results

Table 10 Run 4.3: A-posteriori estimator, absolute and relative errors for the control and state variable for $X = V$, different number ℓ of POD basis functions and for the reference control u^2_{ref}

ℓ	$\|\zeta^\ell\|/\gamma$	$\mathcal{E}^u_{\mathrm{abs}}(\ell)$	$\mathcal{E}^u_{\mathrm{rel}}(\ell)$	$\mathcal{E}^y_{\mathrm{abs}}(\ell)$	$\mathcal{E}^y_{\mathrm{rel}}(\ell)$
1	$1.2\cdot 10^{+1}$	$2.1\cdot 10^{+0}$	$7.20\cdot 10^{-1}$	$1.4\cdot 10^{-1}$	$4.9\cdot 10^{-2}$
5	$6.5\cdot 10^{-1}$	$5.6\cdot 10^{-1}$	$1.88\cdot 10^{-1}$	$2.3\cdot 10^{-2}$	$7.9\cdot 10^{-3}$
15	$2.7\cdot 10^{-2}$	$2.7\cdot 10^{-2}$	$9.0\cdot 10^{-3}$	$7.12\cdot 10^{-4}$	$2.5\cdot 10^{-4}$
20	$7.5\cdot 10^{-3}$	$7.3\cdot 10^{-3}$	$2.4\cdot 10^{-3}$	$1.97\cdot 10^{-4}$	$6.9\cdot 10^{-5}$
40	$3.9\cdot 10^{-4}$	$3.9\cdot 10^{-4}$	$1.3\cdot 10^{-4}$	$8.09\cdot 10^{-6}$	$2.8\cdot 10^{-6}$
60	$8.3\cdot 10^{-5}$	$8.3\cdot 10^{-5}$	$2.8\cdot 10^{-5}$	$1.61\cdot 10^{-6}$	$5.6\cdot 10^{-7}$
70	$3.0\cdot 10^{-5}$	$3.0\cdot 10^{-5}$	$1.0\cdot 10^{-5}$	$6.02\cdot 10^{-7}$	$2.1\cdot 10^{-7}$
90	$3.7\cdot 10^{-6}$	$3.7\cdot 10^{-6}$	$1.3\cdot 10^{-6}$	$1.21\cdot 10^{-7}$	$4.2\cdot 10^{-8}$
130	$3.9\cdot 10^{-8}$	$3.9\cdot 10^{-8}$	$1.3\cdot 10^{-8}$	$4.49\cdot 10^{-9}$	$1.6\cdot 10^{-9}$

Table 11 Run 4.3: Number of differences in restricted node values of the POD suboptimal controls $\bar{u}^\ell$ in comparison to the FE optimal control $\bar{u}^h$ using u^2_{ref} and $X = V$

l	5	10	15	20	30	40	50	60	70	80	90	100	110	120	130
diffua	1021	311	74	28	5	2	0	0	0	0	0	0	0	0	0
diffub	324	147	37	8	0	0	0	0	0	0	0	0	0	0	0

associated with u^1_{ref} in Table 9, the more prudent choice of the reference control and the POD basis leads to more/earlier harmony of restricted values.

References

1. K. Afanasiev, M. Hinze, Adaptive control of a wake flow using proper orthogonal decomposition, in *Lecture Notes in Pure and Applied Mathematics*, vol. 216 (2001), pp. 317–332
2. E. Arian, M. Fahl, E.W. Sachs, Trust-region proper orthogonal decomposition for flow control. Technical Report 2000-25, ICASE, 2000
3. P. Benner, E.S. Quintana-Ortí, Model reduction based on spectral projection methods, in *Reduction of Large-Scale Systems*, ed. by P. Benner, V. Mehrmann, D.C. Sorensen. Lecture Notes in Computational Science and Engineering, vol. 45 (2005), pp. 5–48
4. R. Dautray, J.-L. Lions, in *Evolution Problems I.* Mathematical Analysis and Numerical Methods for Science and Technology, vol. 5 (Springer, Berlin, 1992)
5. L. Dede, Reduced basis method and a posteriori error estimation for parameterized optimal control problems. SIAM J. Sci. Comput. **32**, 997–1019 (2010)
6. M.A. Grepl, M. Kärcher, Reduced basis a posteriori error bounds for parametrized linear-quadratic elliptic optimal control problems. C. R. Acad. Sci. Paris, Ser. I **349**, 873–877 (2011)
7. M. Hintermüller, K. Ito, K. Kunisch, The primal-dual active set strategy as a semi-smooth Newton method. SIAM J. Optim. **13**, 865–888 (2003)
8. M. Hintermüller, I. Kopacka, S. Volkwein, Mesh-independence and preconditioning for solving parabolic control problems with mixed control-state constraints. ESAIM Control Optim. Calc. Var. **15**, 626–652 (2008)

9. M. Hintermüller, M. Ulbrich, A mesh-independence result for semismooth Newton methods. Math. Program., Ser. B **101**, 151–184 (2004)
10. M. Hinze, S. Volkwein, Error estimates for abstract linear-quadratic optimal control problems using proper orthogonal decomposition. Comput. Optim. Appl. **39**, 319–345 (2008)
11. P. Holmes, J.L. Lumley, G. Berkooz, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, Cambridge Monographs on Mechanics (Cambridge University Press, Cambridge, 1996)
12. M. Kahlbacher, S. Volkwein, POD a-posteriori error based inexact SQP method for bilinear elliptic optimal control problems. ESAIM: M2AN **46**, 491–511 (2012)
13. E. Kammann, F. Tröltzsch, S. Volkwein, A method of a-posteriori error estimation with application to proper orthogonal decomposition. ESAIM: Math. Model. Numer. Anal. **47**, 555–581 (2013)
14. K. Kunisch, S. Volkwein, Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics. SIAM J. Numer. Anal. **40**, 492–515 (2002)
15. K. Kunisch, S. Volkwein, Proper orthogonal decomposition for optimality systems. ESAIM: M2AN **42**, 1–23 (2008)
16. J.L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations* (Springer, Berlin, 1971)
17. A. Manzoni, A. Quarteroni, G. Rozza, Shape optimization for viscous flows by reduced basis methods and free-form deformation. Int. J. Numer. Meth. Fluids **70**, 646–670 (2012)
18. B. Noble, *Applied Linear Algebra* (Prentice-Hall, Englewood Cliffs, 1969)
19. J. Nocedal, S.J. Wright, *Numerical Optimization*. Springer Series in Operation Research 2nd edn. (2006)
20. E. Sachs, S. Volkwein, POD-Galerkin approximations in PDE constrained optimization. GAMM-Mitt. **33**, 194–208 (2010)
21. A. Studinger, Numerical analysis of POD a-posteriori error estimates for linear-quadratic optimal control problems. Diploma thesis, Department of Mathematics and Statistics, University of Constance, 2011
22. T. Tonn, K. Urban, S. Volkwein, Comparison of the reduced-basis and POD a-posteriori error estimators for an elliptic linear quadratic optimal control problem. Math. Comput. Model. Dyn. Syst. **17**, 355–369 (2011)
23. F. Tröltzsch, S. Volkwein, POD a-posteriori error estimates for linear-quadratic optimal control problems. Comput. Optim. Appl. **44**, 83–115 (2009)
24. S. Volkwein, Model reduction using proper orthogonal decomposition, Lecture Notes, University of Constance (2011)
25. S. Volkwein, Optimality system POD and a-posteriori error analysis for linear-quadratic problems. Control Cybern. **40**, 1109–1125 (2011)
26. G. Vossen, S. Volkwein, Model reduction techniques with a-posteriori error analysis for linear-quadratic optimal control problems. Numer. Algebra Control Optim. **2**, 465–485 (2012)

Cubature on C^1 Space

Gabriel Turinici

Abstract We explore in this paper cubature formulas over the space of functions having a first continuous derivative, i.e., C^1. We show that known cubature formulas are not optimal in this case and explain what is the origin of the loss of optimality and how to construct optimal ones; to illustrate we give cubature formulas up to (including) order 9.

Keywords Cubature formulas · Stochastic analysis · Chen signature · Chen series · Cubature on infinite dimensional space · Cubature Wiener · Cubature finance

Mathematics Subject Classification (2010) Primary 60H35 · 65D32 · 91G60 · Secondary 65C30 · 65C05

1 Introduction

We consider the following controlled ordinary differential equation (ODE)

$$dx(t) = f\bigl(x(t), u(t)\bigr)dt, \qquad x(0) = x_0, \tag{1.1}$$

where f is supposed as smooth as required with respect to all variables and $u(t)$ a C^1 control that acts on $x(t)$ with $u(0) = u'(0) = 0$. Let T be some final time (which will be set to 1 in all that follows) and denote by $C_0^1([0,T];\mathbb{R})$ the space of u. In order to explicitly mark the dependence of x on u we will also write $x_u(t)$ for the solution of (1.1).

We place ourselves in a situation where many $u(t)$ can be chosen and the average (or any aggregate quantity such as higher order moments, etc.) of some functional of $x(T)$ over all such $u(t)$ is to be computed. Typical frameworks where this is relevant

This work was funded by ANR projects EMAQS number ANR-2011-BS01-017-01 and ISOTACE.

G. Turinici (✉)
CEREMADE, Université Paris Dauphine, Place du Marechal de Lattre de Tassigny,
75016 Paris, France
e-mail: Gabriel.Turinici@dauphine.fr

K. Bredies et al. (eds.), *Control and Optimization with PDE Constraints*,
International Series of Numerical Mathematics 164,
DOI 10.1007/978-3-0348-0631-2_9, © Springer Basel 2013

is in inverse problems where one can chose several controls u, measure the output on the system depending on $x(T)$ and want to identify some parts of the function f by doing this (see [4, 9, 10] for examples).

We need to make precise what average means. Since our primary space for $u(t)$ is $C_0^1([0,T];\mathbb{R})$ a possible way to formalize this average is to consider a one dimensional Brownian motion W_t (we write sometimes, as is usual, the time as index instead of $W(t)$ but this means the same thing) and write the following 3-dimensional stochastic differential equation (SDE):

$$dx(t) = f\big(x(t), u(t)\big)dt, \qquad x(0) = x_0, \tag{1.2}$$

$$du(t) = w(t)dt, \qquad u(0) = 0, \tag{1.3}$$

$$dw(t) = 1dW_t = 1 \circ dW_t, \qquad w(0) = W_0 = 0, \tag{1.4}$$

where the last equality means of course $w(t) = W_t$. The third equality is there only in order to give a formal 3D SDE; the term $\circ dW_t$ signals a Stratonovich formulation (which is the one well adapted to cubature framework because of the Wong–Zakai theorem [15]).

We can now make precise the quantity of interest which is

$$\mathbb{E}F\big(x(T)\big), \tag{1.5}$$

where F is some (smooth enough) real function.

The justification of this formal writing is the following: the Brownian motion selects paths on the (Wiener) space of continuous functions null at the origin on $[0,T]$ denoted $C_0^0([0,T];\mathbb{R})$. Any $C_0^1([0,T];\mathbb{R})$ is the definite primitive of a function in the Wiener space. Thus as realizations of W span the Wiener space, $u(t)$ will span the required space.

Following works on infinite dimensional cubature formulas on Wiener space by [7, 8, 11] (see also [12, 14] for an application of cubature to finance and [3] to SPDE; many other works appeared in the literature on these subjects) we want to approximate the mean in (1.5) by a finite sum

$$\mathbb{E}F\big(x(T)\big) \simeq \sum_{k=1}^{n} \lambda_k F\big(x_{u_k}(T)\big), \tag{1.6}$$

where each u_k corresponds to a given realization ω_k of the Brownian motion W and the corresponding $u_k(t)$ is given as above by

$$u_k(t) = \int_0^t \omega_k(s)ds. \tag{1.7}$$

Such an approximation is called a cubature formula. The question is what weights λ_k and paths ω_k are best for some given n and how good are the approximation properties of such a cubature formula.

A first thought is to use cubature formulas that work on the Wiener space $C_0^0([0,T];\mathbb{R})$ (cf. cited references for the details). As it will be seen in the following this is not necessarily the most efficient choice because of the specific structure of the problem. The purpose of this work is to find optimal cubature formulas for the space $C_0^1([0,T];\mathbb{R})$ up to (including) fourth order.

The plan of the paper is the following: further motivating remarks are the object of Sect. 2 while a quick introduction to cubature formulas on Wiener space is presented in Sect. 3. Preliminary computations are given in Sect. 4 while the actual cubature formulas are given in Sect. 5.

2 Further Remarks and Motivation

Denote $Y = \begin{pmatrix} x \\ u \\ w \end{pmatrix}$ and note that our equation can be written as

$$dY = \begin{pmatrix} f(x(t),u(t)) \\ w \\ 0 \end{pmatrix} dt + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \circ dW_t. \tag{2.1}$$

We note that a different circumstance where the term $\mathbb{E}F(x(T))$ appears is in the forward Kolmogorov (or Fokker–Planck) PDE associated to the time evolution of the density of the SDE (2.1). If we denote by $\rho(t,x,u,w)$ the 3D density it satisfies the following degenerate 3-dimensional, time-dependent PDE [6, 13]:

$$\frac{\partial}{\partial_t}\rho(t,x,u,w) + \frac{\partial}{\partial_x}\big(f(x,u)\rho(t,x,u,w)\big) + \frac{\partial}{\partial_u}\big(w\rho(t,x,u,w)\big)$$
$$- \frac{1}{2}\frac{\partial^2}{\partial_w^2}\rho(t,x,u,w) = 0, \tag{2.2}$$
$$\rho(0,x,u,w) = \delta_{x=x_0}. \tag{2.3}$$

Then since

$$\mathbb{E}F\big(x(T)\big) = \int_{\mathbb{R}_+^3} F(x)\rho(T,x,u,w)dxdudw, \tag{2.4}$$

the method presented here also applies to the evaluation of the right hand side of the equation above. An equivalent formulation, that does not require to work with a Dirac mass, involves a degenerate backward (in time) PDE and can be invoked through the Feynman–Kac formula [6, 13]:

$$\mathbb{E}F\big(x(T)\big) = \mathcal{F}(0,x_0,0,0), \tag{2.5}$$

where

$$\frac{\partial}{\partial_t}\mathcal{F}(t,x,u,w)+f(x,u)\frac{\partial}{\partial_x}\mathcal{F}(t,x,u,w)+w\frac{\partial}{\partial_u}\mathcal{F}(t,x,u,w)$$

$$+\frac{1}{2}\frac{\partial^2}{\partial_w^2}\mathcal{F}(t,x,u,w)=0, \tag{2.6}$$

$$\mathcal{F}(T,x,u,w)=F(x). \tag{2.7}$$

Thus the method presented here can be used to solve degenerate PDEs of type (2.6).

3 Background on Cubature Formulas

We follow [7, 11] and introduce below the principle of computing cubature formulas on the Wiener space. Suppose we want to compute $\mathbb{E}g(Z(T))$ with g a regular function where $Z(t)=(Z_0(t),\dots,Z_d(t))^T\in\mathbb{R}^{d+1}$ solves the SDE

$$dZ=\sum_{\ell=0}^{d+1}\zeta_\ell\bigl(Z(t)\bigr)\circ dB_\ell(t), \tag{3.1}$$

where $B_1(t),\dots,B_d(t)$ are components of a d-dimensional Brownian motion, ζ_l are (generic) smooth functions and we denote $B_0(t)=t$ and set $\zeta_0(\cdot)=1$ (which ensures $Z_0(t)=t$).

If a path $\omega(t)=(\omega_0(t),\dots,\omega_d(t))\in\mathbb{R}^{d+1}$ with $\omega_0(t)=t$ is given and has some regularity one can define $\xi_\omega(t)$ as the solution of the following ODE

$$d\xi_\omega(t)=\sum_{\ell=0}^{d+1}\zeta_\ell\bigl(\xi_\omega(t)\bigr)d\omega_\ell(t). \tag{3.2}$$

Use now stochastic Taylor formulas [6, 13] to write

$$\mathbb{E}g\bigl(Z(T)\bigr)=g\bigl(Z(0)\bigr)+\sum_j a_j(g,\zeta_0,\dots,\zeta_d)\mathbb{E}(P_j)+R, \tag{3.3}$$

where R is a remainder of order higher than a predefined order N; the term $a_j(g,\zeta_0,\dots,\zeta_d)$ is a real (known) functional depending on $g,\zeta_0,\dots,\zeta_d$; and P_j are stochastic polynomials, i.e. integrals of the type

$$\int_0^T\int_0^{s_1}\int_0^{s_2}\dots\int_0^{s_{m-1}}\circ dB_{\alpha_m}(t)\cdots\circ dB_{\alpha_1}(t) \tag{3.4}$$

with $\alpha_p\in\{0,1,\dots,d\}$ for each p. The order of a stochastic polynomial is defined adding 1 for each integral involving $\alpha_j>0$ and 2 for each $\alpha_j=0$.

If the function g is smooth enough and the remainder R does not contain terms of order $\leq N$ a cubature formula of order N

$$\mathbb{E}g\big(X(T)\big) \simeq \sum_{k=1}^{n} \lambda_k \xi_{\omega^k}(T) \tag{3.5}$$

is obtained by requiring that cubature paths ω^k and weights λ_k satisfy for each polynomial P_j as in (3.4):

$$\begin{aligned}&\sum_{k=1}^{n} \lambda_k \bigg(\int_0^T \int_0^{s_1} \int_0^{s_2} \dots \int_0^{s_{m-1}} d\omega^k_{\alpha_m}(t) \dots d\omega^k_{\alpha_1}(t) \bigg) \\ &\quad = \mathbb{E}\bigg(\int_0^T \int_0^{s_1} \int_0^{s_2} \dots \int_0^{s_{m-1}} \circ\, dB_{\alpha_m}(t) \cdots \circ\, dB_{\alpha_1}(t) \bigg).\end{aligned} \tag{3.6}$$

Remark 3.1 We use here the same naming conventions for the order of the cubature scheme as in [11] which is somehow different from the standard numerical analysis practice. As such, a cubature of order "N" will have error of order $O(T^{(N+1)/2})$.

4 Stochastic Taylor Expansion for Averages of Deterministic Functionals over the Class $C_0^1([0, T]; \mathbb{R})$

We will use the following convention: for any function $G(\cdot)$ we denote by $\partial_k G$ the partial derivative of function G with respect to its kth argument. We write the stochastic Taylor formula [13] and iterate:

$$\begin{aligned}\mathbb{E}F\big(x(T)\big) &= \mathbb{E}F\big(x(0)\big) + \mathbb{E}\int_0^T \partial_1 F\big(x(s_1)\big) f\big(x(s_1), u(s_1)\big) ds_1 \\ &= F\big(x(0)\big) + \mathbb{E}\int_0^T \partial_1 F\big(x(s_1)\big) f\big(x(s_1), u(s_1)\big) ds_1.\end{aligned} \tag{4.1}$$

We obtain by iterating:

$$\begin{aligned}\mathbb{E}F\big(x(T)\big) = F\big(x(0)\big) &+ \partial_1 F\big(x(0)\big) f\big(x(0), u(0)\big) \cdot \mathbb{E}\bigg(\int_0^T ds_1\bigg) \\ &+ \mathbb{E}\int_0^T \int_0^{s_1} \partial_1 F\big(x(s_2)\big)(\partial_1 f\big(x(s_2), u(s_2)\big) f\big(x(s_2), u(s_2)\big) \\ &+ \partial_2 f\big(x(s_2), u(s_2)\big) W(s_2)) \\ &+ (\partial_1)^2 F\big(x(s_2)\big) f\big(x(s_2), u(s_2)\big) ds_2 ds_1.\end{aligned} \tag{4.2}$$

The first conclusion that can be drawn from this initial computation is that no first order terms appear and the only second order term in T is $\mathbb{E}(\int_0^T ds_1)$; thus a second order cubature formula (in the sense of the Remark 3.1) has only to satisfy the requirement:

$$\sum_{k=1}^{n} \lambda_k \int_0^T ds_1 = \mathbb{E}\left(\int_0^T ds_1\right) = T, \tag{4.3}$$

i.e.

$$\sum_{k=1}^{n} \lambda_k = 1. \tag{4.4}$$

The important remark here is that many terms are missing among which (we only write terms up to order 3 because the others are more cumbersome to write):

$$\mathbb{E}\left(\int_0^T \circ dW_{s_1}\right), \qquad \mathbb{E}\left(\int_0^T \int_0^{s_1} \circ dW_{s_2} \circ dW_{s_1}\right), \tag{4.5}$$

$$\mathbb{E}\left(\int_0^T \int_0^{s_1} ds_2 \circ dW_{s_1}\right), \qquad \mathbb{E}\left(\int_0^T \int_0^{s_1} \circ dW_{s_2} ds_1\right), \tag{4.6}$$

$$\mathbb{E}\left(\int_0^T \int_0^{s_1} \int_0^{s_2} \circ dW_{s_3} \circ dW_{s_2} \circ dW_{s_1}\right), \tag{4.7}$$

as well as terms of order 4 involving Stratonovich integrals.

It follows that classical cubature formulas derived for fully general equations on Wiener space lose optimality here. The purpose of this work is to explain what are the constraints that optimal cubature formulas satisfy and give examples of optimal weights and paths up to (including) order 9.

Continuing in the same way the enumeration of orders as they appear iterating the integral form of the stochastic Taylor formula we obtain that the following integrals appear:

1. Order 2: term $\mathbb{E}(\int_0^T ds_1)$. The constraint is, as seen above,

$$\sum_{k=1}^{n} \lambda_k = 1. \tag{4.8}$$

2. (Unique) term of order 4: $\mathbb{E}(\int_0^T \int_0^{s_1} ds_2 ds_1)$. There is no new requirement brought by this term.
3. (Unique) term of order 5: $\mathbb{E}(\int_0^T \int_0^{s_1} \int_0^{s_2} \circ dW_{s_3} ds_2 ds_1)$. The requirement is

$$\sum_{k=1}^{n} \lambda_k \left(\int_0^T \int_0^{s_1} \int_0^{s_2} d\omega_k(s_3) ds_2 ds_1\right)$$
$$= \mathbb{E}\left(\int_0^T \int_0^{s_1} \int_0^{s_2} \circ dW_{s_3} ds_2 ds_1\right) = 0. \tag{4.9}$$

We recall that the integral $\int_0^{s_2} d\omega_k(s_3)$ is a Riemann–Stieltjes integral.

4. (Unique) term of order 6: $\mathbb{E}(\int_0^T \int_0^{s_1} \int_0^{s_2} ds_3 ds_2 ds_1)$. There is no new requirement brought by this term.
5. Only two terms of order 7:

$$\mathbb{E}\left(\int_0^T \int_0^{s_1} \int_0^{s_2} \int_0^{s_3} \circ dW_{s_4} ds_3 ds_2 ds_1\right) = 0, \tag{4.10}$$

$$\mathbb{E}\left(\int_0^T \int_0^{s_1} \int_0^{s_2} \int_0^{s_3} ds_4 \circ dW_{s_3} ds_2 ds_1\right) = 0. \tag{4.11}$$

6. Order 8 and higher: all the terms beginning by the terms of order 7 and higher.

5 Cubature Formulas

5.1 *Cubature Formulas of Order* 6

As seen above cubature formulas up to order 4 (included) are somehow trivial. We thus start our list of cubature formulas from order 5. Note that a formula of order 5 is automatically of order 6 too since terms of order 6 do not bring any new requirement (other that the one implied already by the term at order 2).

There are two equations: (4.8) and (4.9). We will use two paths and thus two weights. A natural choice is to use some path ω_1 and $\omega_2 = -\omega_1$ and $\lambda_1 = \lambda_2 = 1/2$. Then the constraints are both satisfied. We obtain for instance a formula of order 6:

$$\lambda_1 = \lambda_2 = 1/2, \qquad \omega_1(t) = t, \qquad \omega_2(t) = -t. \tag{5.1}$$

Note that this is the same as the third order (dimension one) formula from [11].

5.2 *Cubature Formulas of Order* 7

There are two new constraints of order 7. But there constraints are again satisfied if one uses $n = 2$ $\lambda_1 = \lambda_2 = 1/2$ and $\omega_2 = -\omega_1$. Thus e.g. formula (5.1) is also of order 7.

5.3 *Cubature Formulas of Order* 8*: A First Approach*

Two new terms appear that bring new constraints:

$$\mathbb{E}\left(\int_0^T \int_0^{s_1} \int_0^{s_2} \int_0^{s_3} \int_0^{s_4} \circ dW_{s_5} \circ dW_{s_4} ds_3 ds_2 ds_1\right) = \frac{T^4}{48}, \tag{5.2}$$

$$\mathbb{E}\left(\int_0^T\int_0^{s_1}\int_0^{s_2}\int_0^{s_3}\int_0^{s_4}\circ dW_{s_5}ds_4\circ dW_{s_3}ds_2ds_1\right)=0. \tag{5.3}$$

We do not enter here into the specifics of the calculation above (see [1]). In terms of the cubature paths and weights the two new constraints read:

$$\sum_{k=1}^n\lambda_k\left(\int_0^T\int_0^{s_1}\int_0^{s_2}\int_0^{s_3}\int_0^{s_4}d\omega_k(s_5)d\omega_k(s_4)ds_3ds_2ds_1\right)=\frac{T^4}{48}, \tag{5.4}$$

$$\sum_{k=1}^n\lambda_k\left(\int_0^T\int_0^{s_1}\int_0^{s_2}\int_0^{s_3}\int_0^{s_4}d\omega_k(s_5)ds_4d\omega_k(s_3)ds_2ds_1\right)=0. \tag{5.5}$$

Note that the choice $n=2$, $\lambda_1=\lambda_2=1/2$ and $\omega_2=-\omega_1=-t$ does not satisfy these constraints. A first idea is to add two more functions and look for a $n=4$ cubature formula of order 8. In order to build on conclusions from previous lower order we further choose to set

$$\lambda_2=\lambda_1,\qquad \lambda_3=\lambda_4,\omega_2=-\omega_1,\qquad \omega_3=-\omega_4. \tag{5.6}$$

Denoting

$$\begin{aligned}\alpha_k&=\int_0^T\int_0^{s_1}\int_0^{s_2}\int_0^{s_3}\int_0^{s_4}d\omega_k(s_5)d\omega_k(s_4)ds_3ds_2ds_1\\&=\int_0^T\int_0^{s_1}\int_0^{s_2}\int_0^{s_3}\omega_k(s_4)d\omega_k(s_4)ds_3ds_2ds_1\\&=\int_0^T\int_0^{s_1}\int_0^{s_2}\frac{\omega_k^2(s_3)}{2}ds_3ds_2ds_1\end{aligned} \tag{5.7}$$

and

$$\begin{aligned}\beta_k&=\int_0^T\int_0^{s_1}\int_0^{s_2}\int_0^{s_3}\int_0^{s_4}d\omega_k(s_5)ds_4d\omega_k(s_3)ds_2ds_1\\&=\int_0^T\int_0^{s_1}\int_0^{s_2}\int_0^{s_3}\omega_k(s_4)ds_4d\omega_k(s_3)ds_2ds_1\end{aligned} \tag{5.8}$$

we obtain that the following requirements are to be satisfied (for $T=1$):

$$\lambda_1\alpha_1+\left(\frac{1}{2}-\lambda_1\right)\alpha_3=\frac{1}{96}, \tag{5.9}$$

$$\lambda_1\beta_1+\left(\frac{1}{2}-\lambda_1\right)\beta_3=0. \tag{5.10}$$

Let us introduce the parameter $\theta\in\mathbb{R}$ and choose $\omega_1=\theta t=-\omega_2$; we compute and obtain $\alpha_1=\frac{\theta^2}{5!}=\beta_1$. It suffices now to choose a family of functions where to

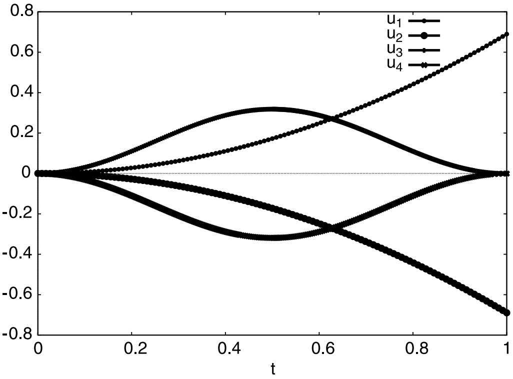


Fig. 1 The four functions u_k for a order 9 quadrature formula

look for ω_3 and its opposite ω_4. Instead of piecewise linear functions as in [11] we propose here oscillatory functions $\omega_3(t) = \sin(\frac{2\pi t}{T}) = -\omega_4$. The unknowns are now θ and λ_1. Note that ω_3 is such that $\int_0^1 \omega_3(t)dt = 0$.

For this choice of ω_3 we obtain (for $T = 1$)

$$\alpha_3 = \frac{8\pi^2 - 3}{192\pi^2}, \qquad \beta_3 = -\frac{8\pi^2 - 21}{96\pi^2}. \tag{5.11}$$

Replacing and solving for θ and λ_1 one obtains:

$$\lambda_1 = \frac{5(2\pi^2 - 21)}{6(2\pi^2 - 15)} \simeq 0.39712223492734, \tag{5.12}$$

$$\theta = \frac{\sqrt{8\pi^2 - 21}}{\sqrt{4\pi^2 - 9}} \simeq 1.378974145172718. \tag{5.13}$$

Note that the natural constraints $\theta^2 > 0$ and $\lambda_1 \in [0, 1/2]$ are satisfied. This is not necessarily the case for other (arbitrary chosen) pairs of functions.

We obtain thus the following integration formula for C_0^1 functions (see also Fig. 1):

$$\lambda_2 = \lambda_1 = \frac{5(2\pi^2 - 21)}{6(2\pi^2 - 15)}, \qquad \lambda_3 = \lambda_4 = \frac{1}{2} - \lambda_1, \qquad \theta = \frac{\sqrt{8\pi^2 - 21}}{\sqrt{4\pi^2 - 9}}, \tag{5.14}$$

$$u_1(t) = \theta\frac{t^2}{2} = -u_2(t), \qquad u_3(t) = \frac{1 - \cos(2\pi t)}{2\pi} = -u_4(t). \tag{5.15}$$

5.4 Minimalistic Cubature Formulas of Order 8 and 9

Another approach to construct a formula of order 8 is to start with $n = 2$ paths and weights but adapt them to satisfy the constraints. We note that any choice:

$$n = 2, \qquad \lambda_1 = \lambda_2 = 1/2, \qquad \omega_2 = -\omega_1 \tag{5.16}$$

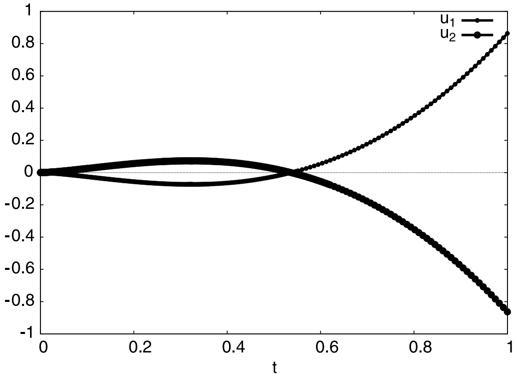


Fig. 2 The functions u_k for a 9th order quadrature formula (5.20)–(5.21)

(now ω_1 is not necessarily t) will automatically satisfy all constraints of odd orders, i.e. involving an odd number of integrations with respect to the paths. The reason is that all such terms have to be zero and are obviously so because are the sum of two contributions, one coming from ω_1 and another, that will have same modulus but opposite sign, from $\omega_2 = -\omega_1$. In particular, if we find a cubature formula of order 8 with two paths that satisfy (5.16) it will also be of order 9.

Thus all that remains to do is to find a function ω_1 which satisfies (5.4) and (5.5) (for $k = 1$). A parametric search as a fractional order polynomial reveals that a suitable solution is:

$$\omega_1(x) = \frac{\sqrt{x}((\sqrt{11}+6)x-3)}{2}. \tag{5.17}$$

A solution can be also found as a piecewise linear function. The function has two linear parts with slope a_1 from 0 to $1/2$ and a_2 from $1/2$ to 1:

$$\omega_1(t) = \frac{(a_2-a_1)|2t-1| + (2a_2+2a_1)t - a_2 + a_1}{4}, \tag{5.18}$$

$$a_1 = -\frac{\sqrt{\sqrt{161}+17}}{2^{\frac{3}{2}}}, \qquad a_2 = -\frac{\sqrt{\sqrt{7}\sqrt{23}+17}(\sqrt{161}-15)}{2^{\frac{5}{2}}}. \tag{5.19}$$

We have thus proved.

Theorem 5.1 *The following choice is a formula of order* 9 *for* $T = 1$ (*see also Fig.* 2):

$$\lambda_1 = \lambda_2 = 1/2, \qquad \omega_1(t) = \frac{\sqrt{t}((\sqrt{11}+6)t-3)}{2}, \qquad \omega_2 = -\omega_1, \tag{5.20}$$

or, in terms of the control u:

$$\lambda_1 = \lambda_2 = 1/2, \qquad u_1(t) = \frac{t^{\frac{3}{2}}((\sqrt{11}+6)t-5)}{5}, \qquad u_2(t) = -u_1(t). \tag{5.21}$$

Same holds for:

$$\omega_1(t) = \frac{(a_2 - a_1)|2t - 1| + (2a_2 + 2a_1)t - a_2 + a_1}{4}, \tag{5.22}$$

$$a_1 = -\frac{\sqrt{\sqrt{161} + 17}}{2^{\frac{3}{2}}}, \qquad a_2 = -\frac{\sqrt{\sqrt{7}\sqrt{23} + 17}(\sqrt{161} - 15)}{2^{\frac{5}{2}}}, \tag{5.23}$$

$$\lambda_1 = \lambda_2 = 1/2, \qquad \omega_2 = -\omega_1, \tag{5.24}$$

or, in terms of the control u:

$$\begin{aligned} u_1(t) &= \frac{((2a_2 - 2a_1)t - a_2 + a_1)|2t - 1| + (4a_2 + 4a_1)t^2}{16} \\ &\quad + \frac{(4a_1 - 4a_2)t + a_2 - a_1}{16}, \\ \lambda_1 &= \lambda_2 = 1/2, \qquad u_2(t) = -u_1(t). \end{aligned} \tag{5.25}$$

Remark 5.2 This methodology to find a cubature formula can be extended to the situation of a multi-dimensional state $x(t)$ and even multi-dimensional control $u(t)$ but the cubature formulas will be different.

Remark 5.3 The point of the paper is that taking into account the special structure of the equation can help to obtain faster cubature formulas. Up to this point f is not depending explicitly on time; if one needs to work with a non-autonomous version of f, the standard treatment is to introduce time as additional variable, but its SDE is very particular ($dt = 1dt$), this may be combined with the above technique (or not...) to propose adapted cubature formulas.

We recall that cubature formulas for arbitrary T are simply obtained by rescaling $\omega_k(t)$ to $\sqrt{T}\omega_k(t/T)$ and $u_k(t)$ to $\sqrt{T^3}\omega_k(t/T)$.

6 Numerical Results

6.1 Linear Setting

To test our implementation against trivial errors we considered first

$$f(x,u) = \alpha x + u, \quad \alpha \in \mathbb{R}, \qquad x(0) = 0, \qquad F(x) = x. \tag{6.1}$$

One can show analytically that $\mathbb{E}F(x(T)) = 0$. The cubature will approximate 0 with $\frac{F(x_{u_1}(T)) + F(x_{u_2}(T))}{2}$. But, by linearity $F(x_{u_2}(T)) = -F(x_{u_1}(T))$ so the approximation is in fact (analytically) exact. The numerical implementation for

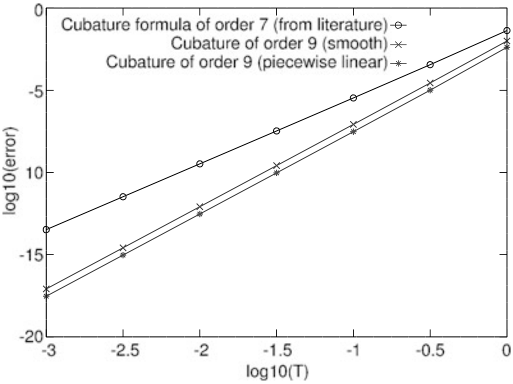


Fig. 3 The error of the 9th order cubature formulas (5.20)–(5.21) and (5.22)–(5.25) for test case (6.2) is plotted in $\log_{10}$–$\log_{10}$ axis. The X axis is $\log_{10}(T)$ and the Y axis is $\log_{10}$ of the error. The resulting plots are *lines* with slope 5. The 7th order formula (5.1) is also plotted, it exhibits a slope of about 4

$\alpha = 0$ and $\alpha = 1$ (results not given here) showed indeed that this is the case i.e. $\frac{F(x_{u_1}(T))+F(x_{u_2}(T))}{2}$ was of the order of the round-off error which in our setting is about 10^{-15}.

6.2 *Nonlinear Setting*

A nonlinear setting was tested next:

$$f(x,u) = x + u^2, \qquad x(0) = 0, \qquad F(x) = x. \tag{6.2}$$

The advantage of this example is that using stochastic expansion we know that

$$\mathbb{E}F\big(x_u(T)\big) = \frac{T^4}{12} + \frac{T^5}{60} + \frac{T^6}{360} + \cdots . \tag{6.3}$$

We tested the cubature formulas of order 9 for different final times in the range $[10^{-3}, 10^0]$. The range was chosen so that the error is not below the round-off. The results in Fig. 3 confirm the theoretical results i.e. the error behaves as $O(T^5)$ for both order 9 formulas. We also tested the formula (5.1) from the literature that uses also only $n = 2$ functions and found $O(T^4)$. As expected, the formulas (5.20)–(5.21) and (5.22)–(5.25) converge faster.

6.3 *Nonlinear Setting Out of the Scope of the Theoretical Result*

Finally, we tested a nonlinear setting taken from [5] which is not of the form (1.1). It involves a 2-dimensional SDE $x = (Y, A)$:

$$dY_t = aY_t dt + bY_t \circ dW_t, \tag{6.4}$$

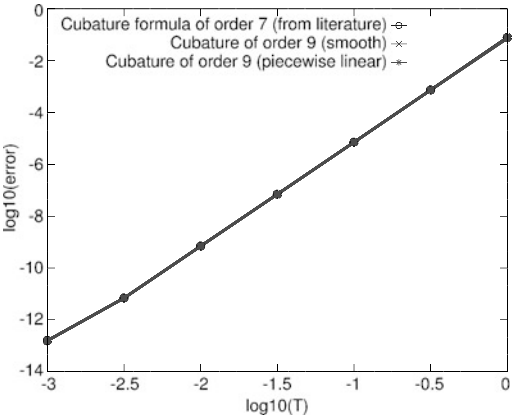


Fig. 4 The error of the 9th order cubature formulas (5.20) and (5.22) and the order 7 formula (5.1) for test case (6.4)–(6.6) is plotted in $\log_{10}$–$\log_{10}$ axis. The X axis is $\log_{10}(T)$ and the Y axis is $\log_{10}$ of the error. The resulting plot is very close to a line of slope 4 for all cubature formulas, the *lines* coincide graphically (but numerical values are different)

$$dA_t = Y_t dt, \qquad a = 0.1, \qquad b = 0.2, \tag{6.5}$$

$$F(x) = A^3. \tag{6.6}$$

Here too, we know the explicit solution of this 2-dimensional SDE

$$Y_t = e^{at+bW_t}, \tag{6.7}$$

$$A_t = \int_0^t e^{as+bW_s} ds. \tag{6.8}$$

The moment $\mathbb{E}(A(T)^3)$ is not trivial to compute and we will not give here its (cumbersome) formula (see instead [16, 17] and also [2] for an elegant way to express it). We tested the same (three) cubature formulas for different final times in the range $[10^{-3}, 10^0]$. The results in Fig. 4 show that for all cubature formulas the error behaves as $O(T^4)$ which says that, from a numerical perspective, all cubatures are of order 7 for test case (6.4)–(6.6). This hints that formulas (5.20) and (5.22) behave at least as well as (5.1) for situations not covered by the theoretical results; recall that all cubatures have the same number of paths $n = 2$.

Acknowledgements Lively and fruitful discussions with Gefry Barad on the topic of the paper are warmly acknowledged.

References

1. F. Baudoin, Stochastic Taylor expansions and heat kernel asymptotics. ESAIM Probab. Stat. **FirstView** (2011)
2. B.J.C. Baxter, R. Brummelhuis, Functionals of exponential Brownian motion and divided differences. J. Comput. Appl. Math. **236**(4), 424–433 (2011)
3. C. Bayer, J. Teichmann, Cubature on Wiener space in infinite dimension. Proc. R. Soc. A, Math. Phys. Eng. Sci. **464**(2097), 2493–2516 (2008)

4. J.M. Geremia, H. Rabitz, Optimal Hamiltonian identification: the synthesis of quantum optimal control and quantum inversion. J. Chem. Phys. **118**(12), 5369–5382 (2003)
5. L. Gergely Gyurkó, T.J. Lyons, Efficient and practical implementations of cubature on Wiener space, in *Stochastic Analysis 2010* (Springer, Heidelberg, 2011), pp. 73–111
6. P.E. Kloeden, E. Platen, *Numerical Solution of Stochastic Differential Equations*. Applications of Mathematics, vol. 23 (Springer, Berlin, 2010). 4th corrected printing
7. S. Kusuoka, Approximation of expectation of diffusion process and mathematical finance, in *Proceedings of the Taniguchi Conference on Mathematics Nara '98*. Adv. Stud. Pure Math., vol. 31 (Mathematical Society of Japan, Tokyo, 2001), pp. 147–165
8. S. Kusuoka, Approximation of expectation of diffusion processes based on Lie algebra and Malliavin calculus, in *Advances in Mathematical Economics*, vol. 6 (Springer, Tokyo, 2004), pp. 69–83
9. C. Le Bris, M. Mirrahimi, H. Rabitz, G. Turinici, Hamiltonian identification for quantum systems: well-posedness and numerical approaches. ESAIM Control Optim. Calc. Var. **13**(02), 378–395 (2007)
10. Z. Leghtas, G. Turinici, H. Rabitz, P. Rouchon, Hamiltonian identification through enhanced observability utilizing quantum control. IEEE Trans. Autom. Control **57**(10), 2679–2683 (2012)
11. T. Lyons, N. Victoir, Cubature on Wiener space. Proc. R. Soc. Lond., Ser. A, Math. Phys. Eng. Sci. **460**(2041), 169–198 (2004)
12. S. Ninomiya, N. Victoir, Weak approximation of stochastic differential equations and application to derivative pricing. Appl. Math. Finance **15**(2), 107–121 (2008)
13. O. Bernt, *Stochastic Differential Equations*, 6th edn. Universitext (Springer, Berlin, 2007)
14. J. Teichmann, Calculating the Greeks by cubature formulae. Proc. R. Soc. A, Math. Phys. Eng. Sci. **462**(2066), 647–670 (2006)
15. E. Wong, M. Zakai, On the convergence of ordinary integrals to stochastic integrals. Ann. Math. Stat. **36**(5), 1560–1564 (1965)
16. M. Yor, On some exponential functionals of Brownian motion. Adv. Appl. Probab. **24**(3), 509–531 (1992)
17. M. Yor, *Exponential Functionals of Brownian Motion and Related Processes*. Springer Finance (Springer, Berlin, 2001)

A Globalized Newton Method for the Optimal Control of Fermionic Systems

Gregory von Winckel

Abstract A computational framework for determining optimal control fields for inducing energy state transitions in systems of several fermions in an infinite potential quantum well is presented. The full multiparticle system is numerically approximated using linear combinations of Slater determinants constructed from nodal trial functions, which leads to diagonalized matrix approximations of variable coefficient terms. First and second order optimality conditions are given for the control and a robust line search is described for computing a local minimizer.

Keywords Optimal control theory · Schrödinger equation · Quantum mechanics · Newton method · Identical particles

Mathematics Subject Classification (2010) Primary 35Q40 · 49M15 · Secondary 49K20 · 90C53

1 Introduction

In recent years there has been growing interest in controlled quantum phenomena by means of external fields. The aim of quantum control is to effect change on a system whose dynamics are governed by the time-dependent Schrödinger equation such that the system reaches a particular configuration. Some applications for quantum control include quantum bits and logic devices, controlled chemical processes, and investigation of fundamental phenomena. Following the initial work of Peirce,

This work was supported by the Austrian Science Fund (FWF) under grant SFB F32 (SFB "Mathematical Optimization and Applications in Biomedical Sciences").

G. von Winckel
Institut für Mathematik und Wissenschaftliches Rechnen, Karl-Franzens-Universität Graz,
36 Heinrichstraße, A-8010 Graz, Austria

Present address:
G. von Winckel (✉)
Center for High Technology Materials, University of New Mexico, 1313 Goddard Street SE,
Albuquerque, NM 87106, USA
e-mail: gregvw@chtm.unm.edu

K. Bredies et al. (eds.), *Control and Optimization with PDE Constraints*,
International Series of Numerical Mathematics 164,
DOI 10.1007/978-3-0348-0631-2_10, © Springer Basel 2013

Daleh, and Rabitz [12], the Lagrangian based optimal control strategy has become prevalent for determining the control field which drives a quantum system closest to a target state at a specified time. Optimal controls are typically computed using either monotonically convergent schemes [10], gradient based schemes such as nonlinear conjugate gradients, BFGS [15], or inexact Newton methods [17].

In practice, multiparticle systems are usually approximated using a many-body approximation such as the Hubbard model, the Born–Oppenheimer approximation, or the Multiconfigurational Time-Dependent Hartree(–Fock) (MCTDH(F)) methods [6, 11], which tends to give a good approximation for large numbers of particles, but may be inaccurate when there are only a few particles. In this current work, we present an efficient discretization and optimal control technique for inducing state transitions in a one-dimensional system of noninteracting or interacting fermions. The basic approach for discretizing the multiple fermion system has been described in detail in [4] and the Krylov–Newton method for a single particle system has been presented in [16]. The current work, however, contains the first application of Newton's method to the multiparticle optimal control problem.

In most quantum control literature where the time-dependent Schrödinger equation (TDSE) is used as an equality constraint, one of the most common approaches is to replace the partial differential equation with a finite dimensional system of ordinary differential equations so there is a two or three level system [1, 2, 5, 13]. This method usually assumes the structure of the Hamiltonian in the eigenfunction basis and further assumes that higher level states play no role in the dynamics. It does, however, capture the basic bilinear structure of the full problem and is attractive since the small systems that result can be solved numerically very quickly.

Alternately, the TDSE may be discretized, for example, with the finite difference method [7], which also gives a finite dimensional, albeit considerably larger, system of equations. This approach is numerically more expensive, but makes direct use of the physical potential and allows for coupling into higher energy states. It may be the case, however, that the discretization may be superfluous in the sense that the method resolves states which do not have a significant occupation probability.

The approach in the current work it to combine both of these ideas. We discretize the multiparticle Hamiltonian directly and compute its eigenvectors which are used as a modal basis for the state. This would be the simplest version of the proper orthogonal decomposition (POD) applied to a symmetric quadratic problem, however, we do not use the term POD in this work as this approach of diagonalizing the Hamiltonian is completely standard practice in quantum mechanics. What is distinct here is that the diagonalization follows a spectral discretization of the full interaction problem and that the basis of eigenvectors is permuted for efficiency of solving the control problem. That is to say, the eigenvectors are ranked in importance to the problem by a heuristic method described below and the state is projected onto the first few selected vectors. The optimal control problem is then solved on this reduced basis. The state space is then augmented by adding the next few most important vectors and the optimization routine is restarted using the computed optimal control from the previous step. This process is repeated until augmenting the state space has no perceptible effect on the cost functional.

The organization of the paper is as follows: in Sect. 2, basic properties of fermionic systems in one dimension are presented. In Sect. 3, we give the space and time discretizations for the multiparticle system. In Sect. 4, the control problem is formulated and the state reduction method is described. Section 5 presents some computed optimal controls for the quantum well containing two, three, and four fermions, and Sect. 6 contains the conclusion and discussion of future work.

2 Multiparticle Systems in One Dimension

To understand the time-dependent Schrödinger equation for multiple interacting fermions, it is advantageous first to consider the two particle case before its generalization to n particles. The TDSE for two fermions is

$$i\partial_t \psi(x_1, x_2, t) = \left\{-\left(\partial_{x_1}^2 + \partial_{x_2}^2\right) + V(x_1, x_2, t)\right\}\psi(x_1, x_2, t). \tag{2.1}$$

The wavefunction $\psi(x_1, x_2, t)$ contains information about both particles and in particular, moreover, following the Born rule, its modulus squared is understood to be a probability density function. The stationary states for the particle system are the solutions to the eigenvalue problem:

$$\left\{-\left(\partial_{x_1}^2 + \partial_{x_2}^2\right) + V(x_1, x_2)\right\}\phi_j(x_1, x_2, t) = \lambda_j \phi(x_1, x_2), \tag{2.2}$$

where the eigenvalue λ_j is the energy. Since these form a complete basis, the time dependent solution can be expanded as a linear combination of the eigenfunctions using time dependent coefficients:

$$\psi(x_1, x_2, t) = \sum_{k=1}^{\infty} c_k(t)\phi_k(x_1, x_2). \tag{2.3}$$

Since the particles are indistinguishable, it is required that this probability function be invariant under exchange of the particles, i.e.,

$$\left|\phi(x_1, x_2)\right|^2 = \left|\phi(x_2, x_1)\right|^2 \quad \Rightarrow \quad \phi(x_2, x_1) = \phi(x_1, x_2)e^{i\theta}. \tag{2.4}$$

This means that $e^{i\theta}$ is the eigenvalue of a permutation operator P where $P\phi(x_1,x_2) = \phi(x_2, x_1) = e^{i\theta}\phi(x_1, x_2)$. Since $P^2 = I$, it follows that for two identical particles that $\phi(x_1, x_2) = \pm\phi(x_2, x_1)$. The Pauli exclusion principle for fermions stipulates that the wavefunction is antisymmetric. In the time independent case, if there is no interaction between the particles, then $V(x_1, x_2) = V_1(x_1) + V_2(x_2)$ and the system is said to be decomposable. The problem is separated into two uncoupled univariate problems by writing the Ansatz

$$\phi(x_1, x_2) = \phi_1(x_1)\phi_2(x_2) - \phi_1(x_2)\phi_1(x_2) = \begin{vmatrix} \phi_1(x_1) & \phi_2(x_1) \\ \phi_1(x_2) & \phi_2(x_2) \end{vmatrix}, \tag{2.5}$$

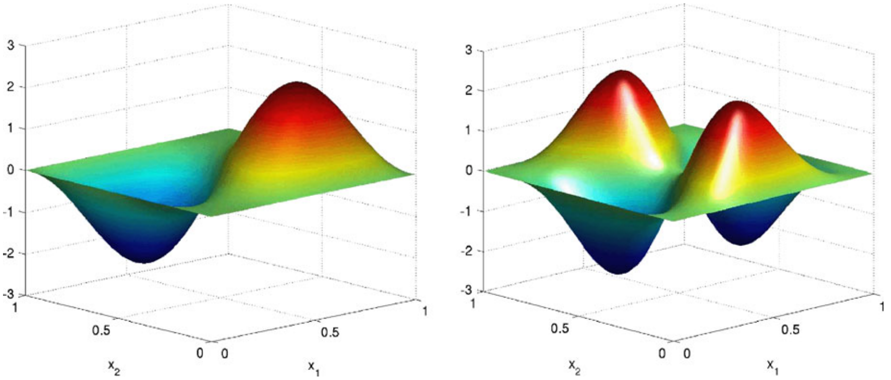


Fig. 1 Two noninteracting fermions: state $|1,2\rangle$ and $|1,3\rangle$

where the determinant on the right hand side is called a Slater determinant. The eigenproblem can be separated into two uncoupled one-dimensional problems

$$\begin{aligned} \left\{-\partial_{x_1}^2 + V_1(x_1)\right\}\phi_{j_1}(x_1) &= \lambda_{j_1}\phi_{j_1}(x_1), \\ \left\{-\partial_{x_2}^2 + V_2(x_2)\right\}\phi_{j_2}(x_2) &= \lambda_{j_2}\phi_{j_2}(x_2). \end{aligned} \tag{2.6}$$

More generally for n non-interacting particles, the stationary states are still Slater determinants of size n:

$$\phi(x_1,\dots,x_n) = \begin{vmatrix} \phi_1(x_1) & \phi_2(x_1) & \cdots & \phi_n(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \cdots & \phi_n(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(x_n) & \phi_2(x_n) & \cdots & \phi_n(x_n) \end{vmatrix}. \tag{2.7}$$

Consider the example of two noninteracting fermions in an infinite potential square quantum well with $x \in [0,1]$. The single particle eigenfunctions and eigenvalues are

$$\phi_j(x) = \sin(\pi j x), \qquad \lambda_j = (j\pi)^2. \tag{2.8}$$

The first two eigenfunctions for the two particle problem are

$$\begin{aligned} |1,2\rangle \equiv \phi_1(x_1,x_2) &= \sin(\pi x_1)\sin(2\pi x_2) - \sin(2\pi x_1)\sin(\pi x_2), \\ |1,3\rangle \equiv \phi_2(x_1,x_2) &= \sin(\pi x_1)\sin(3\pi x_2) - \sin(3\pi x_1)\sin(\pi x_2), \end{aligned} \tag{2.9}$$

and the corresponding eigenvalues are $\lambda_1 = 5\pi^2$ and $\lambda_2 = 10\pi^2$. The first and second states are shown in Fig. 1.

If we add a third noninteracting fermion to the same well, then the first state $|1,2,3\rangle$ is the eigenfunction

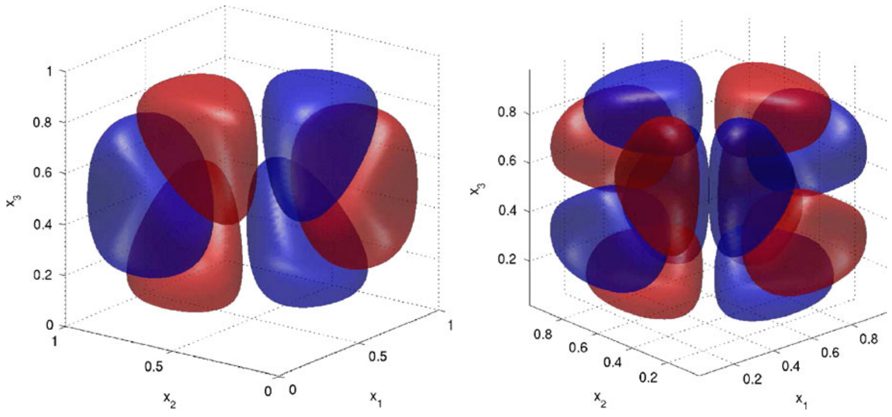


Fig. 2 Three noninteracting fermions: state $|1,2,3\rangle$ and $|1,2,4\rangle$

$$\begin{aligned}\phi_1(x_1,x_2,x_3) = {} & \sin(\pi x_1)\big[\sin(2\pi x_2)\sin(3\pi x_3) - \sin(3\pi x_2)\sin(2\pi x_3)\big] \\ & + \sin(2\pi x_1)\big[\sin(3\pi x_2)\sin(\pi x_3) - \sin(\pi x_2)\sin(3\pi x_3)\big] \\ & + \sin(3\pi x_1)\big[\sin(\pi x_2)\sin(2\pi x_3) - \sin(2\pi x_2)\sin(\pi x_3)\big], \qquad (2.10)\end{aligned}$$

and the first eigenvalue is $\lambda_1 = 14\pi$. The first two states for this system are depicted in Fig. 2 where the red surfaces indicate the level sets where the eigenfunction is equal to half of its maximum value and the blue surfaces correspond to half the minimum value.

In stationary problems with nonzero interaction potential or time-dependent problems, it is no longer the case that the wave function is decomposable in this way and is not a single Slater determinant of one-dimensional functions. In the more general case, however, it is reasonable to write the wavefunction as a linear combination of Slater determinants.

The full multiparticle TDSE that we discretize and use as an equality constraint has the form

$$i\partial_t \psi(\mathbf{x},t) = \left\{-\Delta + \sum_{j=1}^{n}\left(V^c(x_j,t) + \sum_{k>j}^{n} V^i(x_j,x_k)\right)\right\}\psi_i(\mathbf{x},t), \qquad (2.11)$$

where $-1 \le x_j \le 1$ and the external potential experienced by the jth particle, $V^c(x_j,t) = u(t)x_j$ corresponds to a spatially-uniform electric field with time dependent amplitude. The interaction is modeled by a smoothed Coulomb potential

$$V^i(x_j,x_k) = \frac{q}{|x_j - x_k|} \approx \frac{q}{\sqrt{(x_j - x_k)^2 + \delta^2}} \qquad (2.12)$$

with δ as a smoothing factor.

3 Numerical Discretization

The wavefunction is discretized in each spatial dimension using the Legendre–Gauss numerical integration (G-NI) discretization, which is algebraically equivalent to the pseudospectral method, but leads to symmetric matrices. Consequently, all variable coefficient matrices will be diagonal, which is especially important for the interaction potential matrices which would otherwise be full in general.

For a single dimension, the approximation has the form

$$\psi(x) \approx \psi_p(x) = \sum_{k=1}^{p} \hat{\psi}_k \ell_k(x), \qquad \ell_j(x) = \prod_{\substack{k=0\\k\neq j}}^{p+1} \frac{x - x_k}{x_j - x_k}, \tag{3.1}$$

where $j = 0, \ldots, p+1$. To enforce homogeneous Dirichlet conditions, we simply exclude $\ell_0(x)$ and $\ell_{p+1}(x)$. The Legendre–Gauss–Lobatto nodes are implicitly defined by

$$\{x_0, \ldots, x_{p+1}\} = \left\{x | P'_{p+1}(x) = 0\right\} \cup \{\pm 1\}, \tag{3.2}$$

where $P_k(x)$ is the kth Legendre polynomial. From the nodes and Legendre polynomials, we also obtain the corresponding Lobatto weights

$$w_k = \frac{1}{(p+1)(p+2)} \frac{2}{[P_{p+1}(x_k)]^2}. \tag{3.3}$$

Starting with a one-dimensional eigenvalue problem such as in (2.6), expanding the wavefunction in the Lagrange trial basis, multiplying by a Lagrange test function and integrating by parts gives us the weak form of the eigenvalue problem

$$\sum_{k=1}^{p} \left[\left(\ell'_j, \ell'_k\right) + \left(\ell_j, V^c \ell_k\right)\right] \hat{\psi}_k = \lambda \sum_{k=1}^{p} (\ell_j, \ell_k) \hat{\psi}_k, \tag{3.4}$$

where the component matrices here are the stiffness or Laplacian matrix with elements, $\tilde{\mathbf{K}}_{jk} = (\ell'_j, \ell'_k)$, the confining potential matrix $\tilde{\mathbf{V}}^{\mathbf{c}}_{jk} = (\ell_j, V^c \ell_k)$, and the mass matrix $\tilde{\mathbf{M}}_{jk} = (\ell_j, \ell_k)$. The inner products are then computed approximately using Legendre–Gauss–Lobatto quadrature:

$$\tilde{\mathbf{M}}_{jk} = \sum_{i=1}^{p} \ell_j(x_i) \ell_k(x_i) w_i, \qquad \tilde{\mathbf{K}}_{jk} = \sum_{i=1}^{p} \ell'_j(x_i) \ell'_k(x_i) w_i. \tag{3.5}$$

This choice of discretization diagonalizes the mass matrix so that it contains the quadrature weights along the diagonal $\tilde{\mathbf{M}}_{jk} = w_j \delta_{jk}$. Since the quadrature weights are all positive, a trivial Cholesky factorization can be employed:

$$\tilde{\mathbf{M}} = \mathbf{R}^\top \mathbf{R}, \qquad \mathbf{R}_{jk} = \sqrt{w_j} \delta_{jk}. \tag{3.6}$$

Transforming the eigenbasis by $\mathbf{R}$ gives us the algebraically equivalent simple eigenvalue problem in contrast to what was a generalized eigenvalue problem

$$\big[\mathbf{K}+\mathbf{V}^{\mathbf{c}}\big]\hat{\varphi}=\lambda\hat{\varphi}, \quad \mathbf{K}=\mathbf{R}^{-\top}\tilde{\mathbf{K}}\mathbf{R}^{-1},\ \mathbf{V}^{\mathbf{c}}=\mathbf{R}^{-\top}\tilde{\mathbf{V}}^{\mathbf{c}}\mathbf{R}^{-1}. \tag{3.7}$$

Although some integration accuracy is sacrificed to yield diagonal matrix approximations to the variable coefficients, it has been shown that this method is algebraically equivalent to the standard pseudospectral method on the Legendre–Gauss–Lobatto nodes [3], however, in this setting all of the matrices are symmetric.

Following the idea of the Slater determinant formula for the decomposable problem, the multiparticle wave function is discretized using linear combinations of Slater determinants of the one-dimensional Lagrange interpolants.

The n particle trial function φ is a Slater determinant of L^2-normalized Lagrange polynomials and to compute the Galerkin matrices, inner products must be computed involving trial and test functions, the latter being chosen from the same space for symmetry. Supposing we have two Slater determinants $A(x)$ and $B(x)$ such that

$$A(x)=\begin{vmatrix} a_1(x_1) & \cdots & a_n(x_1) \\ \vdots & \ddots & \vdots \\ a_1(x_n) & \cdots & a_n(x_n) \end{vmatrix}, \qquad B(x)=\begin{vmatrix} b_1(x_1) & \cdots & b_n(x_1) \\ \vdots & \ddots & \vdots \\ b_1(x_n) & \cdots & b_n(x_n) \end{vmatrix}, \tag{3.8}$$

then the Löwden rule for Slater inner products [9] states that

$$\big\langle A(x),\, B(x)\big\rangle=\begin{vmatrix} \langle a_1, b_1\rangle & \cdots & \langle a_1, b_n\rangle \\ \vdots & \ddots & \vdots \\ \langle a_n, b_1\rangle & \cdots & \langle a_n, b_n\rangle \end{vmatrix}. \tag{3.9}$$

The total discretized Laplacian or stiffness matrix is $K=K^1+K^2+\cdots+K^n$ where $K^{\nu}_{jk}=\langle \partial_{x_\nu}\varphi_j,\, \partial_{x_\nu}\varphi_k\rangle$. We can write each of these components as

$$K^{\nu}_{jk}=\begin{vmatrix} \delta_{j_1,k_1} & \cdots & \delta_{j_1,k_{\nu-1}} & K_{j_1,k_\nu} & \delta_{j_1,k_{\nu+1}} & \cdots & \delta_{j_1,k_n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \delta_{j_n,k_1} & \cdots & \delta_{j_n,k_{\nu-1}} & K_{j_n,k_\nu} & \delta_{j_n,k_{\nu+1}} & \cdots & \delta_{j_n,k_n} \end{vmatrix}. \tag{3.10}$$

Ordinarily, discretizing an n-dimensional problem with p degrees of freedom per dimension would result in p^n grid points, however, exploiting the antisymmetry relations of the basis functions reduces the degrees of freedom to $N_p=\binom{p}{n}$. All variable coefficient matrices are diagonal, and the multiparticle Laplacian has a sparsity pattern that matches the adjacency matrix of the Johnson graph with the addition of a full diagonal band. The sparsity pattern for the Laplacian when $p=15$, $n=2$ and $p=15$, $n=5$ are displayed in Fig. 3. An efficient method of computing the Laplacian, and variable coefficient matrices, which utilizes the combinatorial structure of the sparsity pattern arising from this discretization to achieve optimal run-time has been recently published [4].

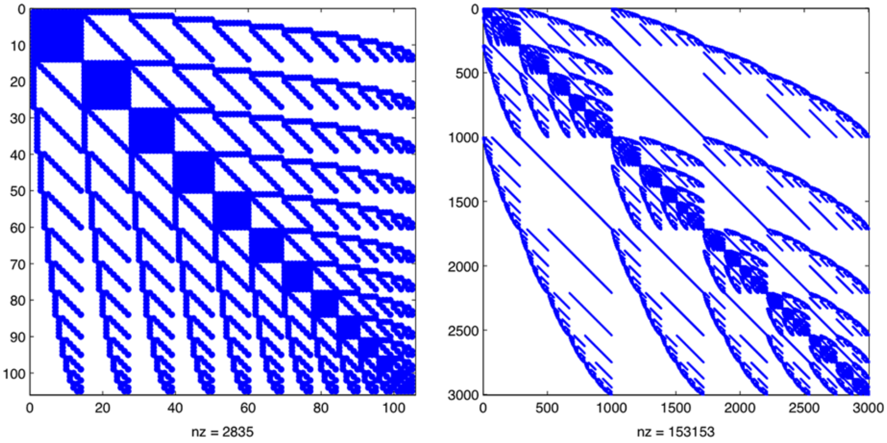


Fig. 3 *Left*: $p = 15$ and $n = 2$. *Right*: $p = 15$ and $n = 5$

After applying the spatial discretization, we obtain a semi-discrete equation of the form

$$i\psi_t = \left\{\mathbf{H}_0 + u(t)\mathbf{V}^{\mathbf{c}}\right\}\psi, \quad \psi \in \mathbb{C}^{N_p}. \tag{3.11}$$

Typically N_p is quite large and it is unnecessary to solve the state equation using all degrees of freedom. Since we are mostly interested in transitions between low lying energy levels, the high level states will usually have extremely low occupancy probability and can be neglected. Instead, we use a reduced order approximation of the state by means of the eigenvalue decomposition. Compute first $N_s \ll N_p$ eigenpairs (Λ, Φ) of stationary Hamiltonian so that

$$\mathbf{H}_0\Phi = \Phi\Lambda, \quad \Phi \in \mathbb{R}^{N_p \times N_s}, \ \Lambda \in \mathbb{R}^{N_s \times N_s}. \tag{3.12}$$

Projecting the state onto the subspace spanned by the computed eigenvectors gives the reduced state equation

$$iy_t = \left\{\Lambda + u(t)\mathbf{X}\right\}y, \quad y \in \mathbb{C}^{N_s}, \ \mathbf{X} = \Phi^\top\mathbf{V}^{\mathbf{c}}\Phi, \tag{3.13}$$

which can be more compactly written as $y_t = i\mathbf{A}(t)y$, where $\mathbf{A}(t) = \Lambda + u(t)\mathbf{X}$.

The Crank–Nicolson method

$$\left(I - \frac{i\delta t}{2}[\mathbf{A}_k + \mathbf{A}_{k-1}]\right)y_k = \left(I + \frac{i\delta t}{2}[\mathbf{A}_k + \mathbf{A}_{k-1}]\right)y_{k-1} \tag{3.14}$$

is one of the more commonly used schemes to numerically integrate the TDSE. It is important to notice, however, that it is only symplectic when the potential is constant over a each time step. This is easily rectified by using the modified Crank–Nicolson method [14], where the control at the endpoints is replaced by the time averaged control over the time step. This approximation retains second order accuracy while making the scheme symplectic. Symplecticity is quite important in quantum control problems as the cost functional can be changed arbitrarily due to numerical loss or gain in the state equation solver otherwise.

4 Control Problem Formulation

Now that the state equation has been discretized in both state and time, the optimal control problem is finite dimensional. The goal is to find the control vector u which is defined on the grid so as to maximize the projection of the state onto the target at the final time. This is formulated as

$$\min_u J(y,\bar{y},u) = 1 - \bar{y}_n^\top \mathbf{P} y_n + \frac{\gamma}{2} u^\top \mathbf{W} u, \tag{4.1}$$

where $\bar{y}$ is the complex conjugate of y, $\mathbf{P}$ is the orthogonal projector onto the target, $0 < \gamma \ll 1$ is a regularization parameter, and $\mathbf{W}$ is the symmetric positive matrix

$$\mathbf{W}_{jk} = \begin{cases} \frac{2\delta t}{3} + \frac{2\epsilon}{\delta t} & \text{if } j = k, \\ \frac{\delta t}{6} - \frac{\epsilon}{\delta t} & \text{if } j = k \pm 1, \\ 0 & \text{otherwise,} \end{cases} \tag{4.2}$$

such that $u^\top \mathbf{W} u$ is a second order approximation of an H^1 type of inner product such as $(u,u) + \epsilon(\dot{u},\dot{u})$, where ϵ is a small positive parameter. Penalizing the derivative of the control enforces the condition that the control go continuously to zero at $t = 0$ and $t = T$. In the numerical experiments, the value $\epsilon = 10^{-3}$ was used.

The fully discretized Schrödinger equation (3.14) provides an equality constraint for every time step, namely that $e_k(y_k, y_{k-1}, u_k, u_{k-1}) = 0$ for $k = 1,\ldots,m$. A Lagrange multiplier is needed for each time step to enforce each equality constraint as well as its complex conjugate. The Lagrangian is

$$L(y,\bar{y},u,\lambda,\bar{\lambda}) = J(y,\bar{y},u) + \sum_{k=1}^{m} \lambda_k^\top e_k + \bar{\lambda}_k^\top \bar{e}_k. \tag{4.3}$$

Taking variations with respect to each of the arguments and setting them to zero gives the first-order optimality conditions. For compact representation, let $B_k = I - \frac{i\delta t}{2}[A_k + A_{k-1}]$. Then we can write the optimality system as

$$\begin{cases} \mathbf{B}_k y_k = \mathbf{B}_k^* y_{k-1}, \quad y_0 \text{ given}, \\ \mathbf{B}_k \lambda_k = \mathbf{B}_{k+1}^* \lambda_{k+1}, \quad \lambda_N = \mathbf{P}\bar{y}_m, \\ \nabla \tilde{J}(u) = Wu - \frac{\delta t}{2} \operatorname{Im}[\xi] = 0, \\ \xi_k = \lambda_k^\top \mathbf{X}(y_k + y_{k-1}) + \lambda_{k+1}^\top \mathbf{X}(y_{k+1} + y_k). \end{cases} \tag{4.4}$$

We can formulate a reduced cost functional by using the fact that the state variable is an implicit function of the control $\tilde{J}(u) = J(y(u),\bar{y}(u),u)$ The control equation in (4.4) expresses the condition that the reduced gradient $\nabla \tilde{J}(u) = 0$.

4.1 Newton's Method

The Lagrangian is not an analytic function of the state and adjoint variables since it also depends on their complex conjugates. Consequently, to compute the Hessian, we use the Wirtinger calculus [8]. In the Wirtinger calculus representation the Hessian is obtained by computing the Jacobian of the complex conjugate of the gradient, so that the Hessian is a complex-valued Hermitian matrix. In particular, L_{ab} really means $\partial_a(\partial_b L)^*$ which is equal to $(L_{ba})^*$. Taking second variations gives rise to the KKT system

$$\begin{pmatrix} L_{yy} & 0 & L_{yu} & 0 & L_{y\bar{\lambda}} \\ 0 & L_{\bar{y}\bar{y}} & L_{\bar{y}u} & L_{\bar{y}\lambda} & 0 \\ L_{uy} & L_{u\bar{y}} & L_{uu} & L_{u\lambda} & L_{u\bar{\lambda}} \\ 0 & L_{\lambda\bar{y}} & L_{\lambda u} & 0 & 0 \\ L_{\bar{\lambda}y} & 0 & L_{\bar{\lambda}u} & 0 & 0 \end{pmatrix} \begin{pmatrix} \delta y \\ \delta\bar{y} \\ \delta u \\ \delta\lambda \\ \delta\bar{\lambda} \end{pmatrix} = - \begin{pmatrix} 0 \\ 0 \\ L_u \\ 0 \\ 0 \end{pmatrix}. \tag{4.5}$$

From the KKT system, we can formally write relationship between the differential change and state and adjoint variables due to a differential change in the control:

$$\begin{aligned} \delta y &= -L_{\bar{\lambda}y}^{-1} L_{\bar{y}u}\delta u, \\ \delta\lambda &= -L_{\bar{y}\lambda}^{-1}[L_{\bar{y}u}\delta u + L_{\bar{y}\bar{y}}\delta\bar{y}]. \end{aligned} \tag{4.6}$$

From this we can write δy and $\delta\lambda$ as the solutions of forced difference equations similar to those for y and λ.

$$\begin{aligned} \mathbf{B}_k\delta y_k &= \mathbf{B}_k^*\delta y_{k-1} + \frac{i\delta t}{2}(\delta u_k + \delta u_{k-1})\mathbf{X}(y_k + y_{k-1}), \\ \mathbf{B}_k\delta\lambda_k &= \mathbf{B}_{k+1}^*\delta\lambda_{k-1} + \frac{i\delta t}{2}(\delta u_k + \delta u_{k+1})\mathbf{X}(\lambda_k + \lambda_{k+1}), \end{aligned} \tag{4.7}$$

where the mth time step for $\delta\lambda$ will also contain an additional term $\mathbf{P}\delta y_m$ on the right hand side.

The action of the reduced Hessian on a test vector δu is

$$\left[\nabla^2\tilde{J}(u)\right]\delta u = L_{uu}\delta u + 2\,\mathrm{Re}[L_{uy}\delta y + L_{u\lambda}\delta\lambda]. \tag{4.8}$$

The Newton search direction can now be computed by iteratively solving the equation

$$\left[\nabla^2\tilde{J}(u)\right]\delta u = -\nabla\tilde{J}(u). \tag{4.9}$$

Since the cost functional is nonconvex, typically even the reduced Hessian is indefinite and the standard conjugate gradient method will not converge. Instead, we use the symmetric LQ (SYMMLQ) method. The computed δu which approximately satisfies (4.9) may not be a descent direction. To handle this possibility, define our

```
Data: Given a descent direction $p$ and the function $f(\alpha) = \tilde{J}(u + \alpha p)$ and
      $f'(\alpha) = p^\top \nabla \tilde{J}(u + \alpha p)$
Compute $\alpha_{\max}$ based on (4.12)
if $\alpha_{\max} > 2$ then
    Evaluate $f(1)$ and $f'(1)$;
    if $\alpha = 1$ satisfies (4.13) then
        $\alpha^* \leftarrow 1$;
    else
        Construct cubic model on [0, 1] and compute its minimum $\alpha_m$;
        Evaluate $f(\alpha_m)$ and $f'(\alpha_m)$
        if $\alpha = \alpha_m$ satisfies SWC then
            $\alpha^* \leftarrow \alpha_m$;
        else
            if $[0, \alpha_m]$ brackets a minimum then
                $\alpha_r \leftarrow \alpha_m$;
            else if [0, 1] brackets a minimum then
                $\alpha_r \leftarrow 1$;
            else
                $\alpha_r \leftarrow \alpha_{\max}$;
            end
            $a^* \leftarrow$ bisect$(0, a_r)$ (Algorithm 2).
        end
    end
else
    $\alpha^* \leftarrow$ bisect$(0, \alpha_{\max})$
end
```

Algorithm 1: Line search algorithm

descent direction as

$$p = \begin{cases} \delta u & \text{if } \delta u^\top \nabla \tilde{J}(u) < 0, \\ -\delta u & \text{if } \delta u^\top \nabla \tilde{J}(u) > 0. \end{cases} \tag{4.10}$$

Of course, should the Newton direction be an ascent direction, it could be discarded in favor of the usual steepest descent direction, however, the scaling of directions produced by solving the Hessian equation tends to be much better. That is to say that the step lengths for sufficient decrease usually remain order 1 instead of order 10^4. Consequently, in our experience, using the sign-flipped Newton direction tends to expedite the line search.

Since the cost functional is nonconvex, a line search strategy (Algorithm 1) is needed to globalize the Newton method. Here we make the observation that the cost functional contains two terms: the physical tracking term $1 - \bar{y}_m^\top \mathbf{P} y_m$ which is uniformly bounded between 0 and 1 and the regularization term which is a pure

```
**Data**: $\alpha_l$ and $\alpha_r$ which bracket a minimum point. $L = \alpha_r - \alpha_l$.
        $f(\alpha) = \tilde{J}(u+\alpha p)$ and $f'(\alpha) = p^\top \nabla \tilde{J}(u+\alpha p)$
while $L > tol$ do
    Compute the midpoint $\alpha_m = \frac{1}{2}(\alpha_l + \alpha_r)$ and evaluate $f(\alpha_m)$ and $f'(\alpha_m)$
    if $\alpha_m$ satisfies (4.13) then
        $\alpha^* \leftarrow \alpha_m$;
    end
    if $f'(\alpha_l) < 0$ and either $f'(\alpha_r) > 0$ or $f(\alpha_r) > f(\alpha_l)$ then
        $\alpha_r \leftarrow \alpha_m$;
    else if $f'(\alpha_l) > 0$ and $f'(\alpha_r) < 0$ or $f(\alpha_r) < f(\alpha_l)$ then
        $\alpha_r \leftarrow \alpha_m$;
    else
        $\alpha_l \leftarrow \alpha_m$;
    end
    $L \leftarrow (\alpha_r - \alpha_l)$
end
```

Algorithm 2: Bisection minimizer

quadratic, consequently once a descent direction p is computed, the reduced cost functional $\tilde{J}(u + \alpha p)$ is an asymptotically quadratic function. This means that the cost functional along the search direction can be bounded from below by the quadratic polynomial

$$\tilde{J}(u+\alpha p) \geq d_2\alpha^2 + d_1\alpha + d_0, \qquad \begin{aligned} d_0 &= \frac{\gamma}{2} u^\top \mathbf{W} u - \tilde{J}(u) \leq 0, \\ d_1 &= \gamma u^\top \mathbf{W} p, \\ d_2 &= \frac{\gamma}{2} p^\top \mathbf{W} p. \end{aligned} \tag{4.11}$$

Since $d_0 \leq 0$, the quadratic equation $d_2\alpha^2 + d_1\alpha + d_0 = 0$ has real roots and we can establish an upper bound on the largest feasible step length α that can still possibly reduce the cost:

$$\alpha_{\max} = \frac{\sqrt{d_1^2 - 4d_0d_2} - d_1}{2d_2}. \tag{4.12}$$

The local minimizer is now guaranteed to satisfy $\alpha^* \in [0, \alpha_{\max}]$.

The strong Wolfe conditions that the step length α must satisfy to give a sufficient decrease in the cost and magnitude of the directional derivative are

$$\begin{aligned} \tilde{J}(u+\alpha p) &\leq \tilde{J}(u) + c_1\alpha p^\top \nabla \tilde{J}(u), \quad && 0 < c_1 \ll 1, \\ \left|p^\top \nabla \tilde{J}(u+\alpha p)\right| &\leq c_2 \left|p^\top \nabla \tilde{J}(u)\right|, \quad && c_1 < c_2 < 1. \end{aligned} \tag{4.13}$$

In the numerical experiments, the values $c_1 = 10^{-4}$ and $c_2 = 0.5$ were used.

4.2 State Model Reduction

Although as a general rule of thumb, the occupancy probability of the lowest energy states are going to be greatest, we can estimate how strong the control potential couples any two states by writing down the state equation in the interaction picture. Given stationary states $\phi_1, \phi_2, \ldots$ and corresponding eigenvalues $\lambda_1, \lambda_2, \ldots$, we can introduce a time dependent change of basis

$$y(t) = \exp(-i\Lambda t)z(t). \tag{4.14}$$

Using this as an Ansatz in our state equation gives the new equation for the transformed state $z(t)$ as

$$\dot{z}(t) = -iu(t)\exp(i\Lambda t)\mathbf{X}\exp(-i\Lambda t)z(t), \tag{4.15}$$

where we can think of the time-dependent similarity transformed matrix $\mathbf{X}$ as being an interaction matrix

$$\tilde{\mathbf{X}}(t) = \exp(i\Lambda t)\mathbf{X}\exp(-i\Lambda t), \tag{4.16}$$

and the specific elements of this matrix are $\tilde{\mathbf{X}}_{jk}\exp(i\omega_{jk}t)$ where $\omega_{jk} = \lambda_j - \lambda_k$. In integral form, the transformed solution at a time t is

$$z(t) = z(0) - i\int_0^t u(\tau)\tilde{\mathbf{X}}(\tau)z(\tau)\,d\tau. \tag{4.17}$$

Of course, $u(t)$ and $z(t)$ are not known in advance; however, $\tilde{\mathbf{X}}(t)$ is known and integration acts as a lowpass filter. In particular, integrating the interaction matrix gives the elements

$$\hat{\mathbf{X}}_{jk}(t) = \int_0^t \tilde{\mathbf{X}}_{jk}(\tau)\,d\tau = \tilde{\mathbf{X}}_{jk}(t)\frac{\exp(i\omega_{jk}t) - 1}{i\omega_{jk}}. \tag{4.18}$$

The magnitude of the $\hat{\mathbf{X}}_{jk}$ gives a rough sense of how strongly the interaction couples state $|j\rangle$ to state $|k\rangle$. Namely, the larger this element is, the more readily we can expect a particle that starts in state $|j\rangle$ to transfer into $|k\rangle$ at some point in time. It stands to reason then, that when considering which basis functions play the most significant role in the dynamics of wavefunction, we should consider not just the difference in eigenvalues, but also the interaction strength.

Once the eigenfunctions are known, the basic idea of the reduced model method is sort the eigenfunctions in decreasing importance to the dynamics, start with only the first few, and then compute the optimal control given that state basis. Once the optimal control is known, the state basis is then enlarged by adding the next few most important states and repeating this process until enlarging the state space no longer has a perceptible effect on the cost functional.

Consider as an example, the problem of two-particle with the control term $u(t)x$, after discretizing and computing the eigenfunctions, we obtain a matrix $\mathbf{X}$ as in

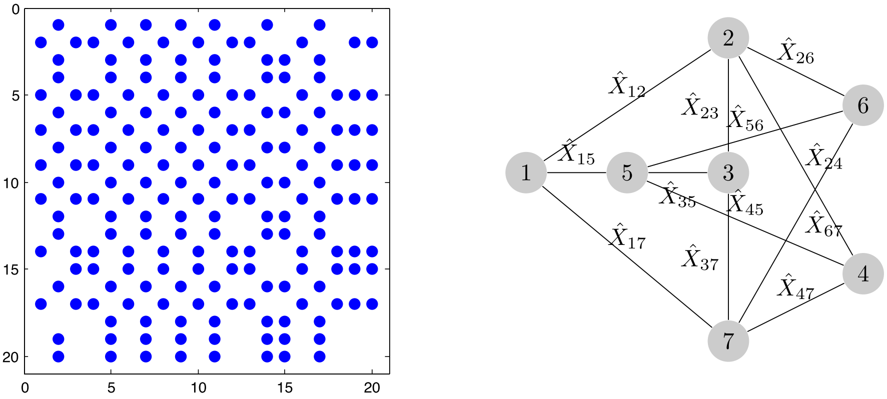


Fig. 4 *Left*: Sparsity pattern for **X** for two particles in a quantum well. *Right*: The connectivity graph for the first seven states

(3.13). The sparsity pattern for this matrix is shown in the left side of Fig. 4. In general, this matrix will be full and this sparsity pattern is a consequence of special parity properties of this problem, however, it is easier to draw the connectivity graph when some states are not directly connected to others. The connectivity graphs shows that, while it is possible to go from any state to any other state, state 1 is not directly connected to state 3, so a particle must go through an intermediate state, such as 2, 5, or 7 first.

The states most strongly directly coupled to the ith state are indicated by the largest elements of the vector

$$r_1 = |e_i + \hat{\mathbf{X}}e_i|, \tag{4.19}$$

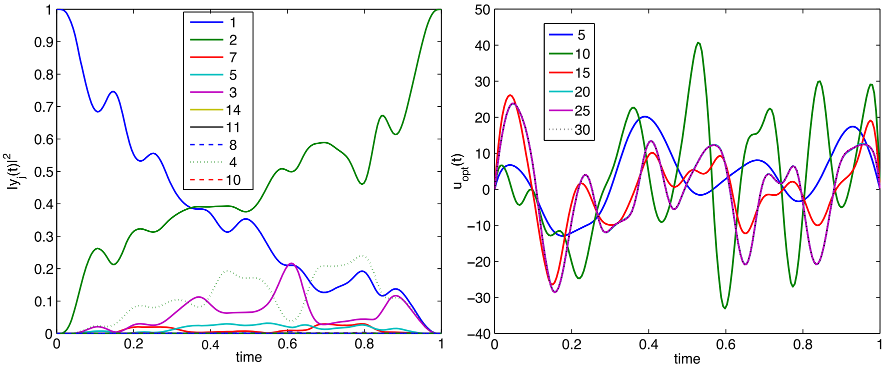
where e_i is the ith canonical vector. By extension, the states most strongly coupled to state i in two steps will be the largest elements of the vector

$$r_2 = \left|e_i + \hat{\mathbf{X}}e_i + \hat{\mathbf{X}}^2 e_i\right|. \tag{4.20}$$

Assuming that any number of intermediate states are allowed, we have the ranking vector

$$r_\infty = \left|\sum_{j=0}^{\infty} \hat{\mathbf{X}}^j e_i\right| = \left|(I - \hat{\mathbf{X}})^{-1} e_i\right|, \tag{4.21}$$

where we are guaranteed that $(I - \hat{\mathbf{X}})^{-1}$ exists since $\hat{\mathbf{X}}$ is skew Hermitian. Recall that the interaction matrix in (4.18) contained an arbitrary phase shift term of $e^{i\omega_{jk}t}$. Although intuitively, this term should be neglected, we did consider the ranking the state importance both with and without this term and found that the reduced model method converges more rapidly when it is neglected.



(a) State occupancy vs. time

(b) Optimal control with refinement

Fig. 5 *Left*: State transition $|1,2\rangle \to |1,3\rangle$ for two interacting particles. *Right*: The computed optimal control for 5, 10 , 15, 20, 25, 30 eigenvectors using the ranking which includes the complex exponential term

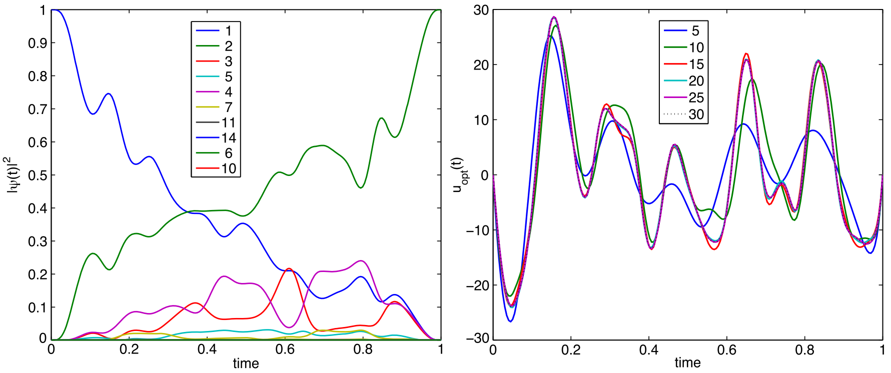
5 Numerical Results

In the test cases, a weak interaction potential (2.12) with unit charge $q = 1$ and smoothing factor $\delta = 0.1$ was used. For both the two and four particle case, 400 uniform time steps were used and the final time was taken to be $T = 1$ for the two particle case and $T = 2$ for the four particle case. For two interacting particles, the order of the eigenstates as ranked by the interaction criteria, excluding the complex exponential term, from strongest to weakest coupling is

$$\{1, 2, 7, 5, 3, 14, 11, 8, 4, 10, 9, 16, 12, 20, 15, 13, 6, 17, 18, 19\}$$

when $N_s = 20$ in (3.12). To compute the control, start with the first five eigenvectors ($N_s = 5$) with indices $\{1, 2, 7, 5, 4\}$ and compute a minimizer by conducting line searches in the Newton search directions. When a local minimizer is obtained, the state space is augmented to include the first ten modes ($N_s = 10$) with indices $\{1, 2, 7, 5, 3, 14, 11, 8, 4, 10\}$. The computed control is no longer a minimizer for the state constraint as there are now allowed transitions into higher states, which increases the objective function. Using the previously computed optimal control from the five dimensional state space as an initial guess, we again compute a sequence of line searches in the Newton directions until we have a new minimizer. This process of augmenting the state space and minimizing until augmenting the space does not noticeably increase the objective function.

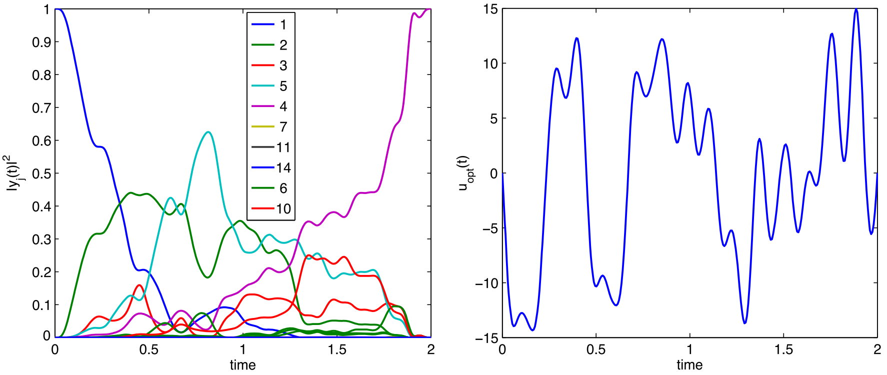
When the complex phase factor in (4.18) is included in determining the ranking, we see that the optimal control in Fig. 5(b) changes significantly as the number of modes N_s increases. When this phase term is excluded, as is the case in Fig. 6(b), the optimal control as a function of N_s stabilizes rapidly. In fact, after $N_s = 25$ there is no significant effect on the cost by further augmenting the space and moreover,



(a) State occupancy vs. time

(b) Optimal control with refinement

Fig. 6 *Left*: State transition $|1,2\rangle \to |1,3\rangle$ for two interacting particles. *Right*: The computed optimal control for $N_s = 5, 10, 15, 20, 25, 30$ modes (eigenvectors) without the complex exponential term. The controls for $N_s = 20$ and $N_s = 30$ appear to be identical



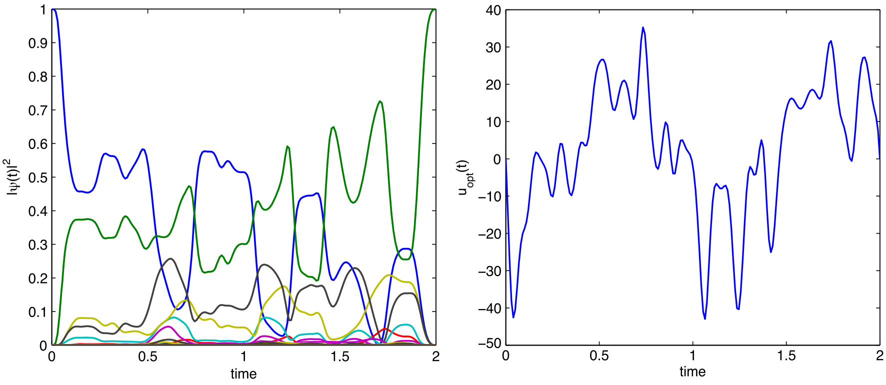
(a) State occupancy vs. time

(b) Optimal control (50 modes)

Fig. 7 *Left*: State transition $|1,2\rangle \to |1,4\rangle$ for two interacting particles. *Right*: The computed optimal control for $N_s = 30$ modes

the optimal control for $N_s = 30$ is not visually distinct from the optimal control for $N_s = 25$. This shows that this ordering strategy is more effective at determining which eigenstates play the most important role in the dynamics. The occupancy of the states shown in Figs. 5(a) and 6(a), is practically identical. It is important to note that the results become effectively discretization independent after twenty or thirty modes are used, which is a significant savings over the full space degrees of freedom $N_p = \binom{15}{2} = 105$.

Thirty modes is not sufficient to resolve the transition from the first state $|1,2\rangle$ to the third state $|1,4\rangle$ as see in Figs. 7(a) and 7(b) due to the significantly higher energy in the control needed to make the transition, as evidenced by the higher fre-



(a) State occupancy vs. time

(b) Optimal control (50 modes)

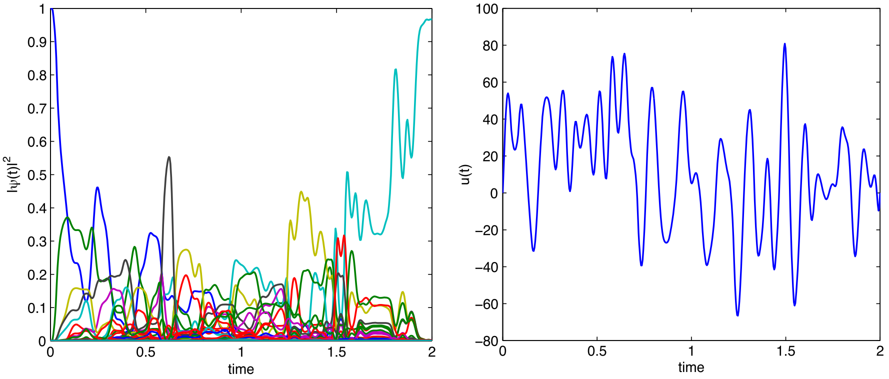
Fig. 8 *Left*: State transition $|1,2,3,4\rangle \rightarrow |1,2,3,5\rangle$ for four interacting particles. *Right*: The computed optimal control for $N_s = 50$ modes

quency terms in the optimal control. Instead fifty modes were needed to adequately resolve the dynamics. More of the states in Fig. 7(a) have a significant occupation probability than was the case with Fig. 6(a).

The problem becomes more challenging as additional particles are added, since the energy spacing of the eigenstates increases considerably. In the four particle system, twenty Lagrange basis functions per particle were needed for the eigenstates to be resolved. This means that before reduction the state has dimension $N_p = \binom{20}{4} = 4845$. However, to compute optimal controls for the transitions between the first state $|1,2,3,4\rangle$ and the second $|1,2,3,5\rangle$ taking the first fifty modes was sufficient. This transition is shown in Fig. 8. Exciting the system to the third state $|1,2,3,6\rangle$ proved quite difficult as at the final time there was only a 97 % probability of finding the particles in the desired state (Fig. 9). We also see that the H^1 norm of the control is becoming quite large and to resolve this problem finer grids will be needed.

In Fig. 10, the reduction of the cost functional and of the L^2 norm of the gradient is shown for the problem of two interacting fermions making the transition from the state $|1,2\rangle$ to the state $|1,3\rangle$. The globalized Newton method is started with a state space of dimension $N_s = 5$, and after 100 iterations, the state space is repeatedly augmented by more sorted eigenfunctions until further augmentation does not appreciably change the cost functional.

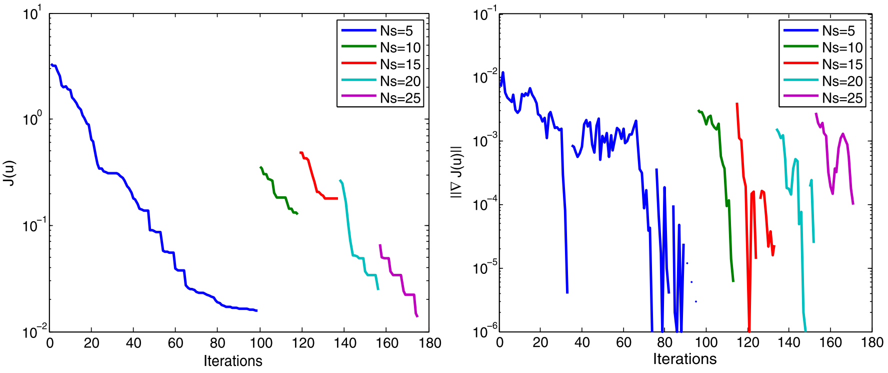
After numerous tests with differing interaction terms, more particles, initial and final states, state space dimension, and time scale, there does not appear to be any completely typical pattern with respect to how often the unit step length of the Newton method yields sufficient decrease and when a line search is needed. It can be said that generally, parameters which make the minimization problem harder are those which increase the energy separation between initial and final states, and this tends to require more line searches.



(a) State occupancy vs. time

(b) Optimal control (50 modes)

Fig. 9 *Left*: State transition $|1,2,3,4\rangle \rightarrow |1,2,3,6\rangle$ for four interacting particles. *Right*: The computed optimal control for $N_s = 50$ modes



(a) Reduction of cost functional

(b) Reduction of gradient

Fig. 10 *Left*: Cost functional value reduction for the state transition $|1,2\rangle \rightarrow |1,3\rangle$ for two interacting particles. *Right*: Reduction of L^2-norm of the gradient

6 Conclusion

We have presented an efficient discretization method and optimization method for controlling energy state transitions for multiple fermions. It was observed that the computational effort of the problem can be reduced by projecting the state onto a suitable reduced basis, which is then augmented as needed to adequately resolve the dynamics. Since optimal control and corresponding state are smooth functions, in future work, higher order symplectic methods will be used to discretize in time so that the number of degrees of freedom in the optimization problem of computing the optimal control can be reduced.

References

1. A. Borzì, G. Stadler, U. Hohenester, Optimal quantum control in nanostructures: theory and application to a generic three-level system. Phys. Rev. A **66**, 053811 (2002)
2. U. Boscain, G. Charlot, J.-P. Gauthier, S. Guerin, H.-R. Jauslin, Optimal control in laser-induced population transfer for two- and three-level quantum systems. J. Math. Phys. **43**(5), 2107–2132 (2002)
3. C.G. Canuto, M. Yousuff Hussaini, A. Quarteroni, T.A. Zang, *Spectral Methods* (Springer, Heidelberg, 2007)
4. C. Clason, G. von Winckel, A general spectral method for the numerical simulation of one-dimensional interacting fermions. Comput. Phys. Commun. **183**(2), 405–417 (2012)
5. M. Grace, C. Brif, H. Rabitz, I.A. Walmsley, R.L. Kosut, D.A. Lidar, Optimal control of quantum gates and suppression of decoherence in a system of interacting two-level particles. J. Phys. B, At. Mol. Opt. Phys. **40**(9), S103 (2007)
6. J. Grond, G. von Winckel, J. Schmiedmayer, U. Hohenester, Optimal control of number squeezing in trapped Bose–Einstein condensates. Phys. Rev. A **80**, 053625 (2009)
7. K. Kime, Finite difference approximation of control via the potential in a 1-d Schrödinger equation. Electron. J. Differ. Equ. **2000**(26), 1–10 (2000)
8. K. Kreutz-Delgado, The complex gradient operator and the CR-calculus, 1–74 (2009). arXiv:0906.4835
9. P.-O. Löwdin, Quantum theory of many-particle systems. I. Physical interpretations by means of density matrices, natural spin-orbitals, and convergence problems in the method of configurational interaction. Phys. Rev. **97**(6), 1474–1489 (1955)
10. Y. Maday, G. Turinici, New formulations of monotonically convergent quantum control algorithms. J. Chem. Phys. **118**(18), 8191–8196 (2003)
11. M. Mundt, D.J. Tannor, Optimal control of interacting particles: a multi-configuration time-dependent Hartree–Fock approach. New J. Phys. **11**(10), 105038 (2009)
12. A.P. Peirce, M.A. Dahleh, H. Rabitz, Optimal control of quantum-mechanical systems: existence, numerical approximation, and applications. Phys. Rev. A **37**, 4950–4964 (1988)
13. D. Sugny, C. Kontz, H.R. Jauslin, Time-optimal control of a two-level dissipative quantum system. Phys. Rev. A **76**(2), 023419 (2007)
14. V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*. Springer Series in Computational Mathematics (Springer, Secaucus, 2006)
15. G. von Winckel, A. Borzì, Computational techniques for a quantum control problem with H^1-cost. Inverse Probl. **24**(3), 034007 (2008)
16. G. von Winckel, A. Borzì, Qucon: a fast Krylov–Newton code for dipole quantum control problems. Comput. Phys. Commun. **181**(12), 2158–2163 (2010)
17. G. von Winckel, A. Borzì, S. Volkwein, A globalized Newton method for the accurate solution of a dipole quantum control problem. SIAM J. Sci. Comput. **31**(6), 4176–4203 (2009)

A Priori Error Estimates for Optimal Control Problems with Constraints on the Gradient of the State on Nonsmooth Polygonal Domains

Winnifried Wollner

Abstract In this article we are concerned with the finite element discretization of optimal control problems subject to a second order elliptic PDE and additional pointwise constraints on the gradient of the state.

We will derive error estimates for the convergence of the cost functional under mesh refinement. Subsequently error estimates for the control and state variable are obtained.

As an intermediate tool we will also analyze a Moreau-Yosida regularized version of the optimal control problem. In particular we will derive convergence rates for the cost functional and the primal variables. To this end we will employ new techniques in estimating the L^{∞}-norm of the feasibility error which could also be used to improve existing estimates in the state constrained case.

Keywords A priori error estimates · Pointwise gradient state constraints · Optimization with PDEs · Finite elements · Nonconvex polygonal domain

Mathematics Subject Classification (2010) 49M25 · 65N12 · 65N30

1 Introduction

We are concerned with an analysis of the discretization error for optimal control problems of second order elliptic equations subject to constraints on the gradient of the state. Such problems have some natural application for instance in cooling processes or structural optimization when high stresses have to be avoided.

Despite these interesting applications first order state constraints have hardly been recognized in mathematics. In the works [3, 4] the case of optimal control of semilinear elliptic equations with pointwise first order state constraints was studied under the assumption that the domain $\Omega \subset \mathbb{R}^n$ possesses a $C^{1,1}$ boundary. In particular, they studied the adjoint equation and derived first order neces-

W. Wollner (✉)
Department of Mathematics, University of Hamburg, Bundesstrasse 55, D-20146 Hamburg, Germany
e-mail: winnifried.wollner@math.uni-hamburg.de

K. Bredies et al. (eds.), *Control and Optimization with PDE Constraints*,
International Series of Numerical Mathematics 164,
DOI 10.1007/978-3-0348-0631-2_11, © Springer Basel 2013

sary optimality conditions. It is immediately clear that their results carry over to the case of a bounded polygonal domain, as long as the linearized state equation (with homogeneous Dirichlet boundary values) defines an isomorphism between $W^{2,t}(\Omega) \cap H_0^1(\Omega)$ and $L^t(\Omega)$ for some $t > n$. However, even for $n = 2$ this requires a convex domain which is usually too restrictive for applications. In a recent publication [20] it was shown that even on nonconvex domains such problems may remain well posed.

In [11] a Moreau-Yosida based framework for PDE-constrained optimization with constraints on the derivative of the state is developed and used to develop a semismooth Newton algorithm. Unfortunately their work does not directly carry over to our problem class, because the presence of corner singularities is contradicting the assumptions made in there article. In [17] an investigation of barrier methods for this problem class is conducted.

When concerned with the discretization of the infinite dimensional problem using finite elements, recent results where obtained in [5, 8, 14]. However in all cases the domain was either smooth or polygonally bounded with sufficiently small interior angles. Concerning adaptive discretization methods we refer to [19] and the recent contribution [10].

The rest of this article is structured as follows. In Sect. 2, we will discuss the problem class under consideration. Then we will consider its discretization in Sect. 2.1. In Sect. 3, we will derive an priori error estimate for a certain semi discretization of the problem. The estimates are essentially the same as those obtained in [5, 8, 14]. Unfortunately for this semi discretization the control has to be chosen orthogonal to certain dual singular functions. Since this is in general not feasible we require further analysis. For this purpose we consider a Moreau-Yosida regularization of the state constraint in Sect. 4. Here, we will derive convergence of both the cost functional and the primal variables depending on the penalization parameter. Parts of the analysis will be similar to the work of [9] but with further complications due to the missing regularity of the control-to-state mapping. We will however employ a new L^∞-estimate for the feasibility violation which could also be used to improve the convergence results obtained in [9, 18] for state constrained problems. For the case without corner singularities one can find similar results obtained simultaneously in [12]. With these preparations we can finally derive the main convergence result, in Sect. 5, for a computationally feasible discretization.

2 Problem Formulation

In what follows, let $\Omega \subset \mathbb{R}^2$ be a bounded polygonal domain. We are concerned with optimization problems governed by a linear elliptic PDE. For simplicity we consider

$$-\Delta u = q \quad \text{in } \Omega, \tag{2.1a}$$

$$u = 0 \quad \text{on } \partial\Omega. \tag{2.1b}$$

It is then clear, that this operator defines an isomorphism $-\Delta\colon V = H^1_0(\Omega) \to H^{-1}(\Omega)$.

Now we let $r > 2$ be a given number and define $Q = L^r(\Omega)$. We are then particularly interested in an optimal control problem of the form

$$\underset{Q\times V}{\text{Minimize}}\ J(q,u) := \frac{1}{2}\left\|u - u^d\right\|^2_{L^2} + \frac{1}{r}\|q\|^r_Q, \tag{2.2a}$$

$$\text{such that } (u,q) \text{ satisfies (2.1a)–(2.1b)}, \tag{2.2b}$$

$$\text{and such that } |\nabla u| \leq 1 \text{ in } \overline{\Omega}. \tag{2.2c}$$

If Ω would be a smooth domain, or a convex polygon, well posedness of (2.2a)–(2.2c) would follow, e.g., from [4]. However, for a general polygon Ω the results do not carry over easily. This is due to the conflicting nature of the constraint $|\nabla u| \leq 1$ and the existence of corner singularities due to the reentrant corners of the domain. This means that for given $q \in Q$ the solution u of (2.1a)–(2.1b) is neither in $C^1(\overline{\Omega})$ nor in $W^{1,\infty}(\Omega)$. Thus the constraint $|\nabla u| \leq 1$ can not be posed easily in this topology. Nonetheless, problem (2.2a)–(2.2c) is well posed, see [20]. The reason for this is that the subspace of controls q that do not give the desired regularity is finite dimensional in our case. In particular (2.2a)–(2.2c) admits a unique solution $(\overline{q},\overline{u}) \in Q \times V$. Moreover there exists a number $t > 2$ depending on the angles in the corners of the domain, such that $\overline{u} \in W^{2,t}(\Omega) \cap V$. Denote the image of $W^{2,t}(\Omega) \cap V$ under $-\Delta$ by I then, again following [20], we have that I is closed in $L^t(\Omega)$. With this preparations we have, in addition, that $\overline{q} \in I \cap Q$.

Since we will utilize knowledge on the value t we will briefly recall how one can calculate t following [20, Lemma 2.2]. In order to determine the value t one utilizes the well known fact, see [7, Theorem 4.4.3.7], that the solution u to (2.1a)–(2.1b) admits a singular expansion

$$u - \sum_{c\in\mathcal{C}} \sum_{\substack{j=1\\ \frac{j\pi}{\omega_c}\neq 1}}^{j<\frac{2\omega_c}{\pi t'}} C_{c,j} s_{c,j}(\rho_c,\theta_c) \in W^{2,t}(\Omega).$$

Here, $\mathcal{C}$ denotes the set of corners of the domain, (ρ_c,θ_c) are polar coordinates with respect to the corner c, $C_{c,j}$ are constants depending on the given right hand side q, and the singular functions $s_{c,j}$ are given by

$$s_{c,j}(\rho_c,\theta_c) = \eta_c(\rho_c)\rho_c^{\frac{j\pi}{\omega_c}} \sin\left(\frac{j\pi}{\omega_c}\theta_c\right)$$

with suitable cutoff functions η_c.

Then in order to assert that $u \in W^{2,t}(\Omega)$ one needs to get $C_{c,j} = 0$. This is where the constraint $|\nabla u| \leq 1$ comes into play. From knowledge that $u \in W^{1,\infty}(\Omega)$ one can infer that $C_{c,j} = 0$ whenever $\frac{j\pi}{\omega_c} < 1$. Now, one can choose $t > 2$ such that the second sum in the expansion only involves indices j such that this is satisfied.

For the exposition of this article it is convenient to assume that Ω has only one reentrant corner v with interior angle $\omega > \pi$. Then from the requirement, $\frac{j\pi}{\omega} < 1$ whenever $j < \frac{2\omega}{\pi t'}$ one can easily calculate

$$\begin{aligned} t &< \frac{\omega}{\pi}\left(\frac{\omega}{\pi} - 1\right)^{-1} && \text{if } \omega \in (\pi, 2\pi), \\ t &< 4 && \text{if } \omega = 2\pi. \end{aligned} \tag{2.3}$$

Further restrictions on t are possible due to convex corners of Ω.

This means that we are able to restate problem (2.2a)–(2.2c) equivalently as follows

$$\underset{Q \cap I \times V}{\text{Minimize}}\ J(q, u) := \frac{1}{2}\left\|u - u^d\right\|_{L^2}^2 + \frac{1}{r}\|q\|_Q^r, \tag{2.4a}$$

$$\text{such that } (u, q) \text{ satisfies (2.1a)–(2.1b)}, \tag{2.4b}$$

$$\text{and such that } |\nabla u| \leq 1 \text{ in } \overline{\Omega}. \tag{2.4c}$$

2.1 Discretization

In a next step we consider the discretization of these problems. To this end we start by discretizing the state equation (2.1a)–(2.1b).

Let $(\mathcal{T}_h)_{h \in (0,1]}$ be a given family of triangulations, consisting of triangles or quadrilaterals which are *affine-equivalent* to their respective reference elements, such that $\operatorname{diam}(T) \leq h$ for all $T \in \mathcal{T}_h$, $h \in (0, 1]$. We assume throughout that the family is quasi-uniform in the sense of [2, Definition 4.4.13], that is, there exists $\rho > 0$ such that, for each $T \in \mathcal{T}_h$ and $h \in (0, 1]$ there exists a ball $B_T \subset T$ such that $\operatorname{diam}(B_T) \geq \rho h$.

We define the discrete state space $V_h \subset V$ as the space of continuous piecewise linear (or bi-linear) functions with respect to the mesh $\mathcal{T}_h$.

We remark that the restrictions we imposed on the family $(\mathcal{T}_h)_{h \in (0,1]}$ ensure that the usual interpolation error results, best approximation results, and inverse estimates hold [2, Sects. 4 and 5].

Finally, we define $\Pi_h : L^1(\Omega) \to V_h$ to be the natural extension of the L^2-projection operator, that is, for $q \in L^1(\Omega)$, we define $\Pi_h u \in V_h$ via

$$(\Pi_h q, \varphi) = (q, \varphi) \quad \forall \varphi \in Q^h. \tag{2.5}$$

It is shown in [6] that Π_h is stable as an operator from $L^p(\Omega)$ to $L^p(\Omega)$, for any $p \in [1, \infty]$, that is, there exist constants c_p, independent of h, such that

$$\|\Pi_h f\|_{L^p} \leq c_p \|f\|_{L^p} \quad \forall f \in L^p(\Omega). \tag{2.6}$$

Now we can discretize the state equation. For fixed $q \in Q$ we search for a solution $u_h \in V_h$ of the following

$$(\nabla u_h, \nabla \varphi_h) = (q, \varphi_h) \quad \forall \varphi_h \in V_h. \tag{2.7}$$

This is already sufficient to obtain a finite dimensional optimization problem. This is due to the fact, that it is sufficient to consider equivalence classes of functions $q, p \in Q$ given by the identification $\Pi_h q = \Pi_h p$ as controls. Then for the minimization of the cost functional in (2.2a)–(2.2c) it is sufficient to take the unique element out of these classes with minimal L^r-Norm. Moreover, due to the first order optimality conditions these elements can be expressed in an explicit way, see, e.g., [13] where this idea was explored first.

In particular, the discretized version of (2.2a)–(2.2c) becomes

$$\underset{Q \times V_h}{\text{Minimize}} \; J(q_h, u_h) := \frac{1}{2} \|u_h - u^d\|_{L^2}^2 + \frac{1}{r} \|q_h\|_Q^r, \tag{2.8a}$$

$$\text{such that } (u_h, q_h) \text{ satisfies (2.7)}, \tag{2.8b}$$

$$\text{and such that } |\nabla u_h| \leq 1 \text{ a.e. in } \overline{\Omega}. \tag{2.8c}$$

In addition, we can also discretize (2.4a)–(2.4c) and get

$$\underset{Q \cap I \times V_h}{\text{Minimize}} \; J(q_h, u_h) := \frac{1}{2} \|u_h - u^d\|_{L^2}^2 + \frac{1}{r} \|q_h\|_Q^r, \tag{2.9a}$$

$$\text{such that } (u_h, q_h) \text{ satisfies (2.7)}, \tag{2.9b}$$

$$\text{and such that } |\nabla u_h| \leq 1 \text{ a.e. in } \overline{\Omega}. \tag{2.9c}$$

Unlike the continuous case the optimal control problems (2.8a)–(2.8c) and (2.9a)–(2.9c) are not equivalent. We will start with an analysis of (2.9a)–(2.9c). This analysis will follow the lines of the arguments used in [14] where some of the arguments have to be refined due to the presence of corner singularities. However, (2.9a)–(2.9c) is not useful in practical computations. This is because the restriction to the controls to lie in I can not be imposed. Hence we will continue our exposition with the analysis of (2.8a)–(2.8c) based upon the results obtained during the analysis of (2.9a)–(2.9c).

Remark 2.1 We remark, that the space I is characterized by a so called dual singular function s_{-1} which is known. However the characterization involves the unknown solution u of (2.1a)–(2.1b). In particular, a function $q \in I$ if and only if with the corresponding solution u to (2.1a)–(2.1b) it holds

$$(q, s_{-1}) + (u, \Delta s_{-1}) = 0.$$

This representation is still of some use in order to calculate the singular coefficients and thereby accelerating convergence of the finite element method for the primal problem, see, e.g., [1]. In order to keep the presentation simple we will not follow such ideas to improve convergence of the discrete problems.

3 Analysis of the Semi Discretization

In this section we will analyze the error between (2.9a)–(2.9c) and (2.2a)–(2.2c) or (2.4a)–(2.4c) respectively.

To this end, we denote the unique solution to (2.2a)–(2.2c) or (2.4a)–(2.4c) by $(\overline{q}, \overline{u})$. The unique solutions to (2.9a)–(2.9c) will be denoted by $(\overline{q}_h^{\perp}, \overline{u}_h^{\perp})$.

Then similar to the proof of [14, Theorem 1] we obtain

Theorem 3.1 *Let* $(\overline{q}, \overline{u}) \in Q \cap I \times W^{2,t}(\Omega) \cap V$ *be the solution to* (2.2a)–(2.2c) *with* $r \geq t > 2$. *Further, let* $(\overline{q}_h^{\perp}, \overline{u}_h^{\perp}) \in Q \cap I \times V_h$ *be the solutions to* (2.9a)–(2.9c). *Then, for any* $\varepsilon > 0$, *there exists a constant* $C > 0$ *independent of* $h \in (0, 1]$ *such that*

$$\left| J(\overline{q}, \overline{u}) - J\left(\overline{q}_h^{\perp}, \overline{u}_h^{\perp}\right) \right| \leq C h^{\beta}$$

where $\beta = 1 - 2/t - \varepsilon$.

Proof We begin our proof by considering the Ritz projection $u_h \in V_h$ of $\overline{u}$ defined by

$$(\nabla u_h, \nabla \varphi_h) = (\overline{q}, \varphi_h) \quad \forall \varphi_h \in V_h.$$

Then, because $\overline{u} \in W^{2,t}(\Omega) \subset W^{1,\infty}$ we have by [15, Theorem 2]

$$\|\nabla \overline{u} - \nabla u_h\|_{\infty} \leq c h^{\beta} \|\overline{u}\|_{C^{1,1-2/t}} \leq c h^{\beta} \|\overline{q}\|_Q. \tag{3.1}$$

Apart from this argument the rest of the proof is the same as in the case of a smooth domain. In particular, with $\tilde{c} \geq c\|\overline{q}\|_Q$, we get that

$$\left(1 - \tilde{c}h^{\beta}\right)|\nabla u_h| \leq \left(1 - \tilde{c}h^{\beta}\right)|\nabla \overline{u}| + \left(1 - \tilde{c}h^{\beta}\right) c h^{\beta} \|\overline{q}\|_Q \leq 1 \quad \text{a.e. in } \Omega.$$

From this we get that

$$(\tilde{q}_h, \tilde{u}_h) = \left(1 - \tilde{c}h^{\beta}\right)(\overline{q}, u_h) \tag{3.2}$$

defines an element $(\tilde{q}_h, \tilde{u}_h) \in Q \cap I \times V_h$ which is feasible for (2.9a)–(2.9c). Now, using the definition of $\tilde{q}_h$ and standard L^2-estimates for $\overline{u} - u_h$, i.e., it is

$$\|\overline{u} - u_h\| \leq c h^{2\pi/\omega} \|\overline{q}\| \leq c h^{\beta} \|\overline{q}\|,$$

it is clear, that

$$\|\overline{q} - \tilde{q}_h\|_Q + \|\overline{u} - \tilde{u}_h\|_2 \leq c h^{\beta} \|\overline{q}\|_Q.$$

Hence, we get

$$\left| J(\overline{q}, \overline{u}) - J(\tilde{q}_h, \tilde{u}_h) \right| \leq c h^{\beta} \|\overline{q}\|_Q$$

from local Lipschitz continuity of J near $(\overline{q}, \overline{u})$. Furthermore we have

$$J\left(\overline{q}_h^{\perp}, \overline{u}_h^{\perp}\right) \leq J(\tilde{q}_h, \tilde{u}_h)$$

because $(\tilde{q}_h, \tilde{u}_h)$ is feasible for (2.9a)–(2.9c). This yields

$$J\left(\overline{q}_h^{\perp}, \overline{u}_h^{\perp}\right) - J(\overline{q}, \overline{u}) \le J(\tilde{q}_h, \tilde{u}_h) - J(\overline{q}, \overline{u}) \le ch^{\beta}.$$

In particular, the sequence $\|\overline{q}_h^{\perp}\|_Q^r \le r J(\overline{q}_h^{\perp}, \overline{u}_h^{\perp}) \le J(\overline{q}, \overline{u}) + ch^{\beta}$ is bounded.

In order to show the reverse inequality, i.e.,

$$-ch^{\beta} \le J\left(\overline{q}_h^{\perp}, \overline{u}_h^{\perp}\right) - J(\overline{q}, \overline{u})$$

we use the same line of arguments. We define for each given solution $(\overline{q}_h^{\perp}, \overline{u}_h^{\perp}) \in Q \cap I \times V_h$ to (2.9a)–(2.9c) a continuous function $u \in V$ using

$$(\nabla u, \nabla \varphi) = \left(\overline{q}_h^{\perp}, \varphi\right) \quad \forall \varphi \in V.$$

Due to the fact that $\overline{q}_h^{\perp} \in Q \cap I$ we have $u \in W^{2,t}(\Omega)$ and hence we get from [15, Theorem 2] that

$$\left\|\nabla \overline{u}_h^{\perp} - \nabla u\right\|_{\infty} \le ch^{\beta} \|u\|_{C^{1,1-2/t}} \le ch^{\beta} \left\|\overline{q}_h^{\perp}\right\|_Q$$

as in (3.1). Now one can continue analog by shifting to obtain a pair $(\hat{q}, \hat{u})$ which is feasible for (2.2a)–(2.2c) such that

$$\left|J\left(\overline{q}_h^{\perp}, \overline{u}_h^{\perp}\right) - J(\hat{q}, \hat{u})\right| \le ch^{\beta}.$$

Note that the constant c is independent of h because $\overline{q}_h^{\perp}$ is bounded independent of h.

This yields the desired lower bound, i.e.,

$$-ch^{\beta} \le J\left(\overline{q}_h^{\perp}, \overline{u}_h^{\perp}\right) - J(\hat{q}, \hat{u}) \le J\left(\overline{q}_h^{\perp}, \overline{u}_h^{\perp}\right) - J(\overline{q}, \overline{u}) \le J(\tilde{q}_h, \tilde{u}_h) - J(\overline{q}, \overline{u}) \le ch^{\beta}$$

and concludes the proof. □

The convergence of the cost functional implies convergence of the primal variables due to strong convexity of J.

Corollary 3.1 *Let $(\overline{q}, \overline{u}) \in Q \cap I \times W^{2,t}(\Omega) \cap V$ be the solution to* (2.2a)–(2.2c) *with $r \ge t > 2$. Further, let $(\overline{q}_h^{\perp}, \overline{u}_h^{\perp}) \in Q \cap I \times V_h$ be the solutions to* (2.9a)–(2.9c). *Then, for any $\varepsilon > 0$, there exists a constant $C > 0$ independent of $h \in (0, 1]$ such that*

$$\left\|\overline{q} - \overline{q}_h^{\perp}\right\|_Q^r + \left\|\overline{u} - \overline{u}_h^{\perp}\right\|^2 \le Ch^{1-2/t-\varepsilon}.$$

Proof The proof is identical to the one for [14, Corollary 1]. To this end let $\tilde{u}_h$ and $\tilde{q}_h$ be given as in the proof of Theorem 3.1. Then by application of Clarkson's inequality to $\overline{u}_h^{\perp} - u^d$ and $\tilde{u}_h - u^d$ and $\overline{q}_h^{\perp}$ and $\tilde{q}_h$ we get

$$\frac{1}{2}\left\|\overline{u}_h^{\perp} - \tilde{u}_h\right\|^2 + \frac{1}{r}\left\|\overline{q}_h^{\perp} - \tilde{q}_h\right\|_Q^r \le \frac{1}{2} J\left(\overline{q}_h^{\perp}, \overline{u}_h^{\perp}\right) + \frac{1}{2} J(\tilde{q}_h, \tilde{u}_h) - J\left(\frac{1}{2} e\right)$$

with $e = (\overline{q}_h^{\perp}, \overline{u}_h^{\perp}) - (\tilde{q}_h, \tilde{u}_h)$. Now since $\frac{1}{2}e$ and $(\tilde{q}_h, \tilde{u}_h)$ are feasible for the discrete problem we get from Theorem 3.1

$$\begin{aligned}\frac{1}{2}\left\|\overline{u}_h^{\perp} - \tilde{u}_h\right\|^2 + \frac{1}{r}\left\|\overline{q}_h^{\perp} - \tilde{q}_h\right\|_Q^r &\leq \frac{1}{2}J(\tilde{q}_h, \tilde{u}_h) - \frac{1}{2}J\left(\overline{q}_h^{\perp}, \overline{u}_h^{\perp}\right)\\ &\leq \frac{1}{2}J(\overline{q}, \overline{u}) - \frac{1}{2}J\left(\overline{q}_h^{\perp}, \overline{u}_h^{\perp}\right) + ch^{\beta} \leq ch^{\beta}.\end{aligned}$$

This shows the assertion using triangle inequality and $\|\overline{u} - \tilde{u}_h\| + \|\overline{q} - \tilde{q}_h\|_Q \leq ch^{\beta}$. □

4 Regularization

Before we come to the analysis of the error between (2.2a)–(2.2c) and (2.8a)–(2.8c), we will need some additional analysis. In particular, we are interested in the following regularized problems for given $\gamma > 0$

$$\begin{aligned}&\underset{Q \times V}{\text{Minimize}}\ J_{\gamma}(q, u) := J(q, u) + \frac{\gamma}{2}\left\|\left(|\nabla u| - 1\right)^{+}\right\|^2,\\ &\text{such that } (u, q) \text{ satisfies (2.1a)–(2.1b).}\end{aligned} \tag{4.1}$$

Similar problems have been analyzed in [11]. Unfortunately their analysis was done under the assumption, that the state equation (2.1a)–(2.1b) defines an isomorphism between $W^{2,t}(\Omega) \cap V$ and $L^t(\Omega)$ which is not the case in our setting. Further we will require bounds on the rate of convergence of the primal variables similar to those obtained in [9] with the improvements made in [12]. Again the arguments are complicated by the fact, that the state equation does not yield sufficient regularity.

We note, that despite these complications one can show convergence of the sequence of minimizers of the problem (4.1) to those of (2.2a)–(2.2c). However, since the convergence is driven by the decay of the L^{∞}-violation of the feasibility and thus, see [12] by the regularity of the solutions $(\overline{q}_{\gamma}, \overline{u}_{\gamma})$ to (4.1) the convergence speed may be dominated by the existence of the corner singularities. As we know that they do not appear in the solution we will apply an additional filter to remove at least parts of the influence of the reentrant corner.

To do so we need to separate the influence of the corner singularities. Hence we define the set $I^{\perp}$ as

$$I^{\perp} = \left\{p \in Q^* = L^{r'}(\Omega) \mid (q, p) = 0\ \forall q \in Q \cap I\right\}$$

where $r' = \frac{r}{r-1}$. The set $I^{\perp}$ is a finite dimensional linear space generated by so called dual singular functions. By the choice of t made in (2.3) it holds $\dim L^{t'}(\Omega)/I = 1$ and thus $\dim I^{\perp} = 1$.

Then we can define the finite dimensional linear space $Q_s \subset Q$ as follows

$$Q_s = \left\{q \in Q \mid \exists p \in I^{\perp}\colon (q, p) \neq 0\right\} \cup \{0\}.$$

This gives the following decomposition of Q as direct sum

$$Q = Q \cap I \oplus Q_s.$$

Let $\{q^\perp\}$ be a basis of $I^\perp$. Then we choose $\{q^s\} \subset Q_s$ as dual basis to $\{q^\perp\}$, i.e., $(q^s, q^\perp) = 1$.

In particular, we can write any element $q \in Q$ as

$$q = q^r + \alpha q^s$$

where $q^r \in Q \cap I$ and $\alpha = \alpha_q = (q, q^\perp) \in \mathbb{R}$ are uniquely determined. In particular $q \in Q \cap I$ if and only if $\alpha = 0$. Corresponding to this relation we can also rewrite any solution u to (2.1a)–(2.1b) with right hand side q as

$$u = u^r + \alpha u^s$$

where u^r and u^2 are given as solutions to (2.1a)–(2.1b) with right hand sides q^r and q^s respectively. Then $u^r \in W^{2,t}(\Omega) \cap V$ and u^s behaves as $r^{\pi/\omega}$ in the vicinity of the reentrant corner. Note that u^s is independent of q.

Then we can state our regularized problem as follows

$$\underset{Q\times V}{\text{Minimize}}\ J_\gamma(q,u) := J(q,u) + \frac{\gamma}{2}\left\|\left(|\nabla u| - 1\right)^+\right\|^2 + \frac{\gamma^2}{2}\left|\left(q, q^\perp\right)\right|^2, \qquad \text{such that } (u,q) \text{ satisfies (2.1a)–(2.1b).} \tag{4.2}$$

We note, that by standard arguments, there exists unique solutions $(\overline{q}_\gamma, \overline{u}_\gamma) \in Q \times V$ to (4.2). Further, let $(\overline{q}_\gamma, \overline{u}_\gamma) \in Q \times V$ be the solution to (2.2a)–(2.2c). Due to the fact, that

$$J(\overline{q}_\gamma, \overline{u}_\gamma) \le J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma) \le J_\gamma(\overline{q}, \overline{u}) = J(\overline{q}, \overline{u})$$

we immediately obtain boundedness of $\|\overline{q}_\gamma\|_Q$. Further, this gives the relation

$$\left\|\left(|\nabla \overline{u}_\gamma| - 1\right)^+\right\|^2 \le C\gamma^{-1}, \qquad |\alpha_\gamma|^2 = \left|\left(\overline{q}_\gamma, q^\perp\right)\right|^2 \le C\gamma^{-2}. \tag{4.3}$$

Our analysis starts with an analysis of the feasibility error.

Lemma 4.1 *Let $(\overline{q}_\gamma, \overline{u}_\gamma) \in Q \times V$ be the solution to (4.2). Further, denote $\overline{q}_\gamma = q^r_\gamma + \alpha_\gamma q^s \in Q \cap I \oplus Q_s$ and $\overline{u}_\gamma = u^r_\gamma + \alpha_\gamma u^s$ the corresponding splitting of the state variable.*

Then there exists a constant c independent of $\gamma \ge 1$ such that

$$\left\|\left(\left|\nabla u^r_\gamma\right| - 1\right)^+\right\|_\infty \le c\gamma^{-1/2(\beta/(1+\beta))},$$

where β is the same as in Theorem 3.1.

Proof To obtain the convergence of u_γ^r in the maximum norm, we define

$$f(x) = \big(|\nabla u_\gamma^r| - 1\big)^+.$$

We remark, that by embedding theorems, we know that $f \in C^{0,\beta}(\Omega)$. Then define

$$\varepsilon_\gamma = \max_{x \in \overline{\Omega}} f(x).$$

An easy computation shows that (for $\gamma \geq 1$)

$$\big\|\big(|\nabla u_\gamma^r| - 1\big)^+\big\|^2 \leq c\big(\big\|\big(|\nabla \overline{u}_\gamma| - 1\big)^+\big\|^2 + |\alpha_\gamma|^2\big) \leq c\gamma^{-1}.$$

We assume w.l.o.g. that $\varepsilon_\gamma > 0$. Then by Hölder continuity of f we get that

$$\begin{aligned} c\gamma^{-1} &\geq \|f\|^2 \\ &\geq \int_{\{f \geq \varepsilon_\gamma/2\}} \big|f(x)\big|^2\,dx \\ &\geq \frac{\varepsilon_\gamma^2}{4} \int_{\{f \geq \varepsilon_\gamma/2\}} dx \\ &\geq \frac{\varepsilon_\gamma^2}{4} c\varepsilon_\gamma^{2/\beta} \\ &\geq c\varepsilon_\gamma^{2+2/\beta}. \end{aligned}$$

Hence by definition

$$\begin{aligned} \big\|\big(|\nabla u_\gamma^r| - 1\big)^+\big\|_\infty &= \varepsilon_\gamma \\ &\leq c\gamma^{-\beta/(2\beta+2)} \end{aligned}$$

which shows the assertion. □

The rate of convergence of $\|(|\nabla u_\gamma^r| - 1)^+\|_\infty$ is the same which could be obtained following the analysis of [9, Lemma 3.1]. Unfortunately this rate also limits our ability to derive convergence estimates for the primal variables. Hence we will spend some effort on improving these results, the techniques employed here have been developed simultaneously in [12]. We will derive them here nonetheless because we will have to face some additional difficulties due to the presence of the corner singularities.

Before doing so, we recall that for a solution $(\overline{q}, \overline{u}) \in Q \times V$ to (2.2a)–(2.2c) there exist $\overline{z} \in L^{t'}(\Omega)$ and $\overline{\mu} \in C^*(\overline{\Omega})$ such that the following necessary optimality conditions hold

$$
\begin{aligned}
(\nabla \overline{u}, \nabla \varphi) &= (\overline{q}, \varphi) && \forall \varphi \in V,\\
(-\Delta \varphi, \overline{z}) &= \big(\overline{u} - u^d, \varphi\big)\\
&\quad + \langle \overline{\mu}, \nabla \overline{u} \cdot \nabla \varphi \rangle_{C^* \times C} && \forall \varphi \in W^{2,t}(\Omega) \cap V,\\
\big(|\overline{q}|^{r-2}\overline{q}, \delta q\big) &= -(\delta q, \overline{z}) && \forall \delta q \in Q \cap I,\\
\langle \overline{\mu}, \varphi \rangle_{C^* \times C} &\le 0 && \forall \varphi \in C(\overline{\Omega}), \varphi \le 0,\\
\big\langle \overline{\mu}, |\nabla \overline{u}| - 1 \big\rangle_{C^* \times C} &= 0,
\end{aligned}
\tag{4.4}
$$

see [20, Theorem 3.3]. Further, by standard arguments for a solution $(\overline{q}_\gamma, \overline{u}_\gamma) \in Q \times V$ to (4.2) there exist $\overline{z}_\gamma \in V$ and $\overline{\mu}_\gamma \in L^2(\Omega)$ such that the following holds

$$
\begin{aligned}
(\nabla \overline{u}_\gamma, \nabla \varphi) &= (\overline{q}_\gamma, \varphi) && \forall \varphi \in V,\\
(\nabla \varphi, \nabla \overline{z}_\gamma) &= \big(\overline{u}_\gamma - u^d, \varphi\big) + (\overline{\mu}_\gamma, \nabla \overline{u}_\gamma \cdot \nabla \varphi) && \forall \varphi \in V,\\
\big(|\overline{q}_\gamma|^{r-2}\overline{q}_\gamma, \delta q\big) &= -(\delta q, \overline{z}_\gamma) - \gamma^2 \alpha_\gamma \big(\delta q, q^\perp\big) && \forall \delta q \in Q,\\
\overline{\mu}_\gamma &= \frac{\gamma}{|\nabla \overline{u}_\gamma|} \big(|\nabla \overline{u}_\gamma| - 1\big)^+
\end{aligned}
\tag{4.5}
$$

compare [11].

Lemma 4.2 *Let $(\overline{q}_\gamma, \overline{u}_\gamma) \in Q \times V$ be the solution to (4.2).*

Then there exists a constant c independent of γ such that it holds

$$
\gamma \big\| \big(|\nabla u_\gamma| - 1\big)^+ \big\|_1 \le c.
$$

Proof We obtain from (4.5) by testing the third equation with $\delta q = \overline{q}_\gamma$ and then using the state and adjoint equation that

$$
\begin{aligned}
\|\overline{q}_\gamma\|_Q^r + \gamma^2 |\alpha_\gamma|^2 &= -(\overline{q}_\gamma, \overline{z}_\gamma)\\
&= (-\nabla \overline{u}_\gamma, \nabla \overline{z}_\gamma)\\
&= -\big(\overline{u}_\gamma - u^d, \overline{u}_\gamma\big) - \gamma \Big(\big(|\nabla \overline{u}_\gamma| - 1\big)^+, |\nabla \overline{u}_\gamma|\Big).
\end{aligned}
\tag{4.6}
$$

Now we obtain that $\|\overline{q}_\gamma\|_Q$, $\|\overline{u}_\gamma\|$, and $\gamma^2 |\alpha_\gamma|^2$ are bounded independent of γ because

$$
0 \le J(\overline{q}_\gamma, \overline{u}_\gamma) \le J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma) \le J(\overline{q}, \overline{u}).
$$

Note that $|\nabla \overline{u}_\gamma| \ge 1$ if $(|\nabla \overline{u}_\gamma| - 1)^+ \ne 0$. Hence we get from (4.6) that

$$
\begin{aligned}
\gamma \big\| \big(|\nabla \overline{u}_\gamma| - 1\big)^+ \big\|_1 &\le \gamma \Big(\big(|\nabla \overline{u}_\gamma| - 1\big)^+, |\nabla \overline{u}_\gamma|\Big)\\
&= -\|\overline{q}_\gamma\|_Q^r - \big(\overline{u}_\gamma - u^d, \overline{u}_\gamma\big) - \gamma^2 |\alpha_\gamma|^2\\
&\le c.
\end{aligned}
$$

□

With these preparations we can derive an improved L^∞ estimate.

Lemma 4.3 *Let $(\overline{q}_\gamma, \overline{u}_\gamma) \in Q \times V$ be the solution to (4.2). Further, denote $\overline{q}_\gamma = q^r + \alpha_\gamma q^s \in Q \cap I \oplus Q_s$ and $\overline{u}_\gamma = u^r_\gamma + \alpha_\gamma u^s$ the corresponding splitting of the state variable. Then there exists a constant c independent of $\gamma \geq 1$ such that*

$$\big\|\big(\big|\nabla u^r_\gamma\big| - 1\big)^+\big\|_\infty \leq c\gamma^{-(\beta/(\beta+2))},$$

where β is the same as in Theorem 3.1.

Proof The proof is analog to the one for Lemma 4.1.

To obtain the convergence of u^r_γ in the maximum norm, we define

$$f(x) = \big(\big|\nabla u^r_\gamma\big| - 1\big)^+.$$

We remark, that by embedding theorems, we know that $f \in C^{0,\beta}(\Omega)$. Then define

$$\varepsilon_\gamma = \max_{x \in \overline{\Omega}} f(x).$$

An easy computation shows that

$$\big|\big(\big|\nabla u^r_\gamma\big| - 1\big)^+\big| \leq \big|\big(|\nabla \overline{u}_\gamma| - 1\big)^+\big| + |\alpha_\gamma|\big|\nabla u^s\big|$$

and hence by Lemma 4.2

$$\big\|\big(\big|\nabla u^r_\gamma\big| - 1\big)^+\big\|_1 \leq \big\|\big(|\nabla u_\gamma| - 1\big)^+\big\|_1 + |\alpha_\gamma|\big\|\nabla u^s\big\|_1 \leq c\gamma^{-1}.$$

We assume w.l.o.g. that $\varepsilon_\gamma > 0$. Then by Hölder continuity of f we get that

$$\big|f(x) - f(y)\big| \leq c\|x - y\|^\beta$$

and hence if for some $x^* \in \overline{\Omega}$ it holds $f(x^*) = \varepsilon_\gamma = \max_{x \in \overline{\Omega}} f(x)$ then we have that $f(y) \geq \varepsilon_\gamma/2$ if $c\|x - y\|^\beta \leq \varepsilon_\gamma/2$. This gives

$$\begin{aligned} c\gamma^{-1} &\geq \|f\|_1 \\ &\geq \int_{\{f \geq \varepsilon_\gamma/2\}} \big|f(x)\big|\,dx \\ &\geq \frac{\varepsilon_\gamma}{2} \int_{\{f \geq \varepsilon_\gamma/2\}} dx \\ &\geq \frac{\varepsilon_\gamma}{2} c\varepsilon_\gamma^{2/\beta} \\ &\geq c\varepsilon_\gamma^{1+2/\beta}. \end{aligned} \tag{4.7}$$

Hence by definition

$$\begin{aligned}\big\|\big(|\nabla u_\gamma^r| - 1\big)^+\big\|_\infty &= \varepsilon_\gamma\\ &\le c\gamma^{-(\beta/(\beta+2))}\end{aligned}$$

which shows the assertion. □

Remark 4.1 We remark, that usually the estimate in (4.7) is too pessimistic. For instance, if $\overline{\mu}$ has support on a curve in $\overline{\Omega}$ it is reasonable to assume that in fact

$$\int_{\{f\ge\varepsilon_\gamma/2\}} dx \ge c\varepsilon_\gamma^{1/\beta}$$

yielding the improved rate

$$\big\|\big(|\nabla u_\gamma^r| - 1\big)^+\big\|_\infty \le c\gamma^{\frac{-\beta}{\beta+1}}.$$

Moreover, if $\overline{\mu}$ has a volume contribution, then the set on which the maximum is attained may even be independent of the Hölder continuity, i.e.,

$$\int_{\{f\ge\varepsilon_\gamma/2\}} dx \ge c$$

then yielding the rate

$$\big\|\big(|\nabla u_\gamma^r| - 1\big)^+\big\|_\infty \le c\gamma^{-1}.$$

For more details we refer to the forthcoming publication [12].

We remark that based upon these preparations one can derive estimates for the primal variables following the ideas of [9, Theorem 2.1] with some modifications due to the presence of corner singularities.

Lemma 4.4 *Let* $(\overline{q}_\gamma, \overline{u}_\gamma) \in Q \times V$ *be the solution to* (4.2) *and* $(\overline{q}, \overline{u}) \in Q \times V$ *be the solution to* (2.2a)–(2.2c)*. Further, denote* $\overline{q}_\gamma = q^r + \alpha_\gamma q^s \in Q \cap I \oplus Q_s$ *and* $\overline{u}_\gamma = u_\gamma^r + \alpha_\gamma u^s$ *the corresponding splitting of the state variable.*

Then, the following estimate holds:

$$\begin{aligned}\|\overline{q}_\gamma - \overline{q}\|_Q^r &+ \big\|u_\gamma^r - \overline{u}\big\|^2 + \frac{\gamma}{2}\big\|\big(|\nabla \overline{u}_\gamma|^2 - 1\big)^+\big\|^2\\ &\le \big\langle\big(\big|\nabla u_\gamma^r\big|^2 - 1\big)^+, \overline{\mu}\big\rangle_{C,C^*} + |\alpha_\gamma|\big(\big|\big(\overline{u} - u^d, u^s\big)\big| + \big|\big(|\overline{q}|^{r-2}\overline{q}, q^s\big)\big|\big)\\ &\le C\big(\big\|\big(\big|\nabla u_\gamma^r\big| - 1\big)^+\big\|_\infty + |\alpha_\gamma|\big).\end{aligned}$$

Proof First we remark, that for any r there exist a constant $c > 0$ such that

$$c\|f - g\|_Q^r \le \big(|f|^{r-2}f - |g|^{r-2}g, f - g\big)$$

holds for any $f, g \in L^r(\Omega) = Q$.

This gives in combination with the necessary optimality conditions (4.4) and (4.5)

$$\begin{aligned}c\|\overline{q}_\gamma-\overline{q}\|_Q^r &\le \big(|\overline{q}_\gamma|^{r-2}\overline{q}_\gamma-|\overline{q}|^{r-2}\overline{q},\overline{q}_\gamma-\overline{q}\big)\\&=-(\overline{z}_\gamma,\overline{q}_\gamma-\overline{q})+\big(\overline{z},q_\gamma^r-q\big)-\alpha_\gamma\big(|\overline{q}|^{r-2}\overline{q},q^s\big)-\gamma^2\alpha_\gamma\big(\overline{q}_\gamma-\overline{q},q^\perp\big)\\&\le-(\overline{z}_\gamma,\overline{q}_\gamma-\overline{q})+\big(\overline{z},q_\gamma^r-q\big)-\alpha_\gamma\big(|\overline{q}|^{r-2}\overline{q},q^s\big)+\gamma^2\alpha_\gamma\big(\overline{q},q^\perp\big).\end{aligned}$$

Now noting that $(\overline{q},q^\perp)=0$ we conclude with the necessary optimality conditions (4.4) and (4.5) that

$$\begin{aligned}c\|\overline{q}_\gamma-\overline{q}\|_Q^r &\le-(\overline{z}_\gamma,\overline{q}_\gamma-\overline{q})+\big(\overline{z},q_\gamma^r-q\big)-\alpha_\gamma\big(|\overline{q}|^{r-2}\overline{q},q^s\big)\\&=(\overline{z}_\gamma,\Delta\overline{u}_\gamma-\Delta\overline{u})-\big(\overline{z},\Delta u_\gamma^r-\Delta\overline{u}\big)-\alpha_\gamma\big(|\overline{q}|^{r-2}\overline{q},q^s\big)\\&=-\big(\overline{u}_\gamma-u^d,\overline{u}_\gamma-\overline{u}\big)-\gamma\big(\nabla\overline{u}_\gamma/|\nabla\overline{u}_\gamma|\big(|\nabla\overline{u}_\gamma|-1\big)^+,\nabla\overline{u}_\gamma-\nabla\overline{u}\big)\\&\quad+\big(\overline{u}-u^d,u_\gamma^r-\overline{u}\big)+\big\langle\overline{\mu},\nabla\overline{u}\big(\nabla u_\gamma^r-\nabla\overline{u}\big)\big\rangle-\alpha_\gamma\big(|\overline{q}|^{r-2}\overline{q},q^s\big)\\&=-\|\overline{u}_\gamma-\overline{u}\|^2-\alpha_\gamma\big(\overline{u}-u^d,u^s\big)-\alpha_\gamma\big(|\overline{q}|^{r-2}\overline{q},q^s\big)\\&\quad-\gamma\big(\nabla\overline{u}_\gamma/|\nabla\overline{u}_\gamma|\big(|\nabla\overline{u}_\gamma|-1\big)^+,\nabla\overline{u}_\gamma-\nabla\overline{u}\big)+\big\langle\overline{\mu},\nabla\overline{u}\big(\nabla u_\gamma^r-\nabla\overline{u}\big)\big\rangle.\end{aligned}$$

To proceed we need to rewrite the last two summands on the right hand side. To do so, we note that both $\gamma(|\nabla\overline{u}_\gamma|-1)^+$ and μ are positive, and hence it is sufficient to estimate the arguments. This yields

$$\begin{aligned}-\nabla\overline{u}_\gamma(\nabla\overline{u}_\gamma-\nabla\overline{u})/|\nabla\overline{u}_\gamma| &=\big(-|\nabla\overline{u}_\gamma|^2+\nabla\overline{u}_\gamma\nabla\overline{u}\big)/|\nabla\overline{u}_\gamma|\\&\le\Big(-|\nabla\overline{u}_\gamma|^2+\frac{1}{2}\big(|\nabla\overline{u}_\gamma|^2+|\nabla\overline{u}|^2\big)\Big)/|\nabla\overline{u}_\gamma|\\&=\frac{1}{2}\big(|\nabla\overline{u}|^2-|\nabla\overline{u}_\gamma|^2\big)/|\nabla\overline{u}_\gamma|\\&\le\frac{1}{2}\big(1-|\nabla\overline{u}_\gamma|^2\big)/|\nabla\overline{u}_\gamma|\\&\le\frac{1}{2}\big(1/|\nabla\overline{u}_\gamma|-|\nabla\overline{u}_\gamma|\big).\end{aligned}$$

Now, noting that $|\nabla\overline{u}_\gamma|^{-1}<1$ on the set $\{|\nabla\overline{u}_\gamma|-1>0\}$ we conclude that on this set

$$-\nabla\overline{u}_\gamma(\nabla\overline{u}_\gamma-\nabla\overline{u})/|\nabla\overline{u}_\gamma|\le\frac{1}{2}\big(1-|\nabla\overline{u}_\gamma|\big).$$

Similarly one gets

$$\nabla\overline{u}\big(\nabla u_\gamma^r - \nabla\overline{u}\big) \le \frac{1}{2}\big(|\nabla u_\gamma^r|^2 - 1\big)^+.$$

Hence the first of the inequalities follows.

The second of the inequalities follows immediately by noting, that $\overline{u}$, $\overline{q}$, u^s, and q^s are independent of γ and that

$$\big|\nabla u_\gamma^r\big|^2 - 1 = \big(|\nabla u_\gamma^r| + 1\big)\big(|\nabla u_\gamma^r| - 1\big) \le c\big(|\nabla u_\gamma^r| - 1\big). \qquad \square$$

We note that Lemma 4.4 combined with Lemma 4.3 immediately gives a bound on the convergence of the primal variables. Additionally one could use the estimate of Lemma 4.1 in a bootstrapping argument to obtain better convergence orders than those derived there. However the results obtained when following this argument are not better than what we obtained in Lemma 4.3. We obtain the following convergence result.

Corollary 4.1 *Let* $(\overline{q}_\gamma, \overline{u}_\gamma) \in Q \times V$ *be the solution to* (4.2) *and* $(\overline{q}, \overline{u}) \in Q \times V$ *be the solution to* (2.2a)–(2.2c).

Then the following estimate holds

$$\|\overline{q}_\gamma - \overline{q}\|_Q^r + \|\overline{u}_\gamma - \overline{u}\|^2 \le c\gamma^{\frac{-\beta}{\beta+2}}.$$

4.1 Convergence Rates for the Cost Functional

Unfortunately, for our later analysis we will require rates of convergence for the cost functionals. Clearly, we get from local Lipschitz continuity of J in combination with Corollary 4.1 that

$$\big|J(\overline{q}_\gamma, \overline{u}_\gamma) - J(\overline{q}, \overline{u})\big| \le c\gamma^{\frac{-\beta}{r(\beta+2)}}.$$

However, as we want to use the difference of the cost functionals to bound the error in the primal variables this is not sufficient. Therefore we will spend some additional effort on the derivation of convergence rates of the cost functional.

Theorem 4.1 *Let* $(\overline{q}_\gamma, \overline{u}_\gamma) \in Q \times V$ *be the solution to* (4.2) *and* $(\overline{q}, \overline{u}) \in Q \times V$ *be the solution to* (2.2a)–(2.2c). *Further, denote* $\overline{q}_\gamma = q^r + \alpha_\gamma q^s \in Q \cap I \oplus Q_s$ *and* $\overline{u}_\gamma = u_\gamma^r + \alpha_\gamma u^s$ *the corresponding splitting of the state variable.*

Assume that

$$\big\|\big(|\nabla u_\gamma^r| - 1\big)^+\big\|_\infty \le c\gamma^{-\theta}$$

then

$$0 \le J(\overline{q}, \overline{u}) - J(\overline{q}_\gamma, \overline{u}_\gamma) \le c\gamma^{-\theta}.$$

Proof Assume that $(1 - c\gamma^{-\theta}) > 0$. Define $\tilde{q}_\gamma = (1 - c\gamma^{-\theta})q_\gamma^r$. Now, denote the corresponding solution to (2.1a)–(2.1b) by $\tilde{u}_\gamma$. Then it holds by assumption that

$$|\nabla \tilde{u}_\gamma| = \big(1 - c\gamma^{-\theta}\big)\big|\nabla u_\gamma^r\big| \le \big(1 - c\gamma^{-\theta}\big)\big(1 + c\gamma^{-\theta}\big) < 1.$$

In particular $(\tilde{q}_\gamma, \tilde{u}_\gamma)$ is feasible for (2.2a)–(2.2c) and hence by local Lipschitz-continuity of J it follows

$$J(\overline{q}, \overline{u}) \le J(\tilde{q}_\gamma, \tilde{u}_\gamma) \le J(\overline{q}_\gamma, \overline{u}_\gamma) + c\gamma^{-\theta}. \qquad \square$$

Finally, we remark that a uniform convexity property holds for the function J_γ.

Lemma 4.5 *The functional J_γ is uniformly convex in the sense that*

$$\frac{1}{2}\|u_1 - u_2\|^2 + \frac{1}{r}\|q_1 - q_2\|_Q^r + J_\gamma\left(\frac{1}{2}(w_1 + w_2)\right) \le \frac{1}{2}J_\gamma(w_1) + \frac{1}{2}J_\gamma(w_2)$$

holds for all $w_1 = (q_1, u_1) \in Q \times V$ and $w_2 = (q_2, u_2) \in Q \times V$.

Proof We note that the stated uniform convexity holds for the cost functional J by application of Clarkson's inequality to $u_1 - u^d$ and $u_2 - u^d$ as well as q_1 and q_2. Hence it remains to show that

$$\left\|\left(\frac{1}{2}|\nabla u_1 + \nabla u_2| - 1\right)^+\right\|^2 \le \frac{1}{2}\big\|\big(|\nabla u_1| - 1\big)^+\big\|^2 + \frac{1}{2}\big\|\big(|\nabla u_2| - 1\big)^+\big\|^2.$$

This is clear, as the integral is monotone, the map $x \mapsto \max(0, x)^2 \colon \mathbb{R} \to \mathbb{R}$ is monotone increasing and $x \to |x| - 1 \colon \mathbb{R}^2 \to \mathbb{R}$ is convex. Similarly we have

$$\frac{1}{4}\big|(q_1 + q_2, q^\perp)\big|^2 \le \frac{1}{2}\big(\big|(q_1, q^\perp)\big|^2 + \big|(q_2, q^\perp)\big|^2\big). \qquad \square$$

Then combination of Theorem 4.1 and Lemma 4.5 yield the same rates of convergence that we obtained in Corollary 4.1.

Corollary 4.2 *Let $(\overline{q}_\gamma, \overline{u}_\gamma) \in Q \times V$ be the solution to (4.2) and $(\overline{q}, \overline{u}) \in Q \times V$ be the solution to (2.2a)–(2.2c).*

Then the following estimate holds

$$\|\overline{q}_\gamma - \overline{q}\|_Q^r + \|\overline{u}_\gamma - \overline{u}\|^2 \le c\gamma^{\frac{-\beta}{\beta+2}}.$$

Proof By Lemma 4.5 we obtain

$$\frac{1}{2}\|\overline{u}_\gamma - \overline{u}\|^2 + \frac{1}{r}\|\overline{q}_\gamma - \overline{q}\|_Q^r \le -J_\gamma\left(\frac{1}{2}(\overline{q}_\gamma + \overline{q}, \overline{u}_\gamma + \overline{u})\right) + \frac{1}{2}J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma) + \frac{1}{2}J_\gamma(\overline{q}, \overline{u})$$

$$\begin{aligned}&\leq \frac{1}{2}J_\gamma(\overline{q},\overline{u}) - \frac{1}{2}J_\gamma(\overline{q}_\gamma,\overline{u}_\gamma)\\ &\leq \frac{1}{2}J(\overline{q},\overline{u}) - \frac{1}{2}J(\overline{q}_\gamma,\overline{u}_\gamma).\end{aligned}$$

This shows the assertion using of Theorem 4.1 and Lemma 4.3. □

Remark 4.2 We comment shortly on the influence of Remark 4.1. Given the comment there the speed of convergence in both Theorem 4.1 as well as in Corollary 4.2 will enhance to

$$J(\overline{q},\overline{u}) - J(\overline{q}_\gamma,\overline{u}_\gamma) \leq c\gamma^{\frac{-\beta}{\beta+1}}$$

in the presence of a line measure in $\overline{\mu}$ and

$$J(\overline{q},\overline{u}) - J(\overline{q}_\gamma,\overline{u}_\gamma) \leq c\gamma^{-1}$$

in the presence of a volume measure in $\overline{\mu}$.

5 Analysis of the Full Discretization

In this section we will analyze the error between (2.8a)–(2.8c) and (2.2a)–(2.2c) or (2.4a)–(2.4c) respectively.

To this end, we denote the unique solution to (2.2a)–(2.2c) or (2.4a)–(2.4c) by $(\overline{q},\overline{u})$. The unique solutions to (2.8a)–(2.8c) will be denoted by $(\overline{q}_h,\overline{u}_h)$.

In contrast to the previous section we can no longer consider the solution $u \in V$ of

$$(\nabla u, \nabla\varphi) = (\overline{q}_h,\varphi) \quad \forall \varphi \in V$$

in order to show that the lower bound

$$-ch^\beta \leq J(\overline{q}_h,\overline{u}_h) - J(\overline{q},\overline{u})$$

holds true. This is because the solution u defined above is no longer an element of $W^{2,t}(\Omega)$. However, from Theorem 3.1 we immediately get

$$J(\overline{q}_h,\overline{u}_h) - J(\overline{q},\overline{u}) \leq J\left(\overline{q}_h^\perp,\overline{u}_h^\perp\right) - J(\overline{q},\overline{u}) \leq ch^\beta$$

because $Q \supset Q \cap I$. In particular, the solutions $\overline{q}_h$ are uniformly bounded.

Before we come to the analysis of the convergence speed, we will start with some preliminary results. First, we will show convergence $\overline{q}_h \to \overline{q}$ and $\overline{u}_h \to \overline{u}$. With these preparations we will compute the distance between $\overline{q}_h$ and I. Then finally, we can obtain the desired convergence rates.

Theorem 5.1 *Let $(\overline{q}_h,\overline{u}_h)$ be the unique solution to (2.8a)–(2.8c) and denote $\overline{q}_h = q_h^r + \alpha_h q^s$ with $\alpha_h = (\overline{q}_h, q^\perp)$. Then $\overline{q}_h \rightharpoonup \overline{q}$ in Q and $\overline{u}_h \to \overline{u}$ in $H_0^1(\Omega)$ where $(\overline{q},\overline{u})$ are the unique solution to (2.2a)–(2.2c). In particular, it holds $\alpha_h \to 0$.*

Proof As already remarked, we have from Theorem 3.1 and $Q \subset Q \cap I$ that

$$J(\overline{q}_h,\overline{u}_h) \le J\big(\overline{q}_h^{\perp},\overline{u}_h^{\perp}\big) \le J(\overline{q},\overline{u}) + ch^{\beta}.$$

This shows that $\|\overline{q}_h\|_Q$ is bounded. Hence there exists a weakly convergent subsequence, denoted again by $\overline{q}_h$, with limit q_0. Due to the compact embedding $L^2(\Omega) \subset H^{-1}(\Omega)$ a subsequence $\overline{q}_h$ converges strongly in $H^{-1}(\Omega)$ and hence $\overline{u}_h$ converges strongly in $H_0^1(\Omega)$ to a limit u_0. Now, for any $\varphi \in H_0^1(\Omega)$ there exists a sequence $\varphi_h \in V_h$ with $\varphi_h \to \varphi$ because $\bigcup_{h>0} V_h$ is dense in $H_0^1(\Omega)$. Thus we have

$$(\nabla u_0,\nabla\varphi) \leftarrow (\nabla\overline{u}_h,\nabla\varphi_h) = (\overline{q}_h,\varphi_h) \to (q_0,\varphi).$$

To proceed, we note that the sequence $|\nabla\overline{u}_h|$ converges strongly in L^2 and hence, again selecting a subsequence, pointwise almost everywhere. Now $\|\nabla\overline{u}_h\|_\infty \le 1$ which shows $\|\nabla u_0\|_\infty \le 1$.

In particular, (q_0,u_0) are feasible for (2.2a)–(2.2c). From weak lower semicontinuity of J we deduce

$$J(q_0,u_0) \le \liminf_{h\to 0} J(\overline{q}_h,\overline{u}_h) \le J(\overline{q},\overline{u}).$$

This shows $q_0 = \overline{q}$ and $u_0 = \overline{u}$. Moreover, since the limit is unique the whole sequence converges.

Finally, because $\overline{q}_h \rightharpoonup \overline{q} \in Q \cap I$ we obtain

$$\alpha_h = \big(\overline{q}_h,q^{\perp}\big) \to \big(\overline{q},q^{\perp}\big) = 0. \qquad \square$$

In a next step we try to obtain a convergence rate for the singular coefficient α_h. To do so, we consider the following problems where q_h^r is given as in Theorem 5.1. We search $u^r, u^s \in V$ and $u_h^r, u_h^s \in V_h$ which solve

$$\big(\nabla u^r,\nabla\varphi\big) = \big(q_h^r,\varphi\big) \quad \forall\varphi \in V, \tag{5.1}$$

$$\big(\nabla u_h^r,\nabla\varphi_h\big) = \big(q_h^r,\varphi_h\big) \quad \forall\varphi_h \in V_h, \tag{5.2}$$

$$\big(\nabla u^s,\nabla\varphi\big) = \big(q^s,\varphi\big) \quad \forall\varphi \in V, \tag{5.3}$$

$$\big(\nabla u_h^s,\nabla\varphi_h\big) = \big(q^s,\varphi_h\big) \quad \forall\varphi_h \in V_h. \tag{5.4}$$

Lemma 5.1 *Let $(\overline{q}_h,\overline{u}_h)$ be the unique solution to (2.8a)–(2.8c) and denote $\overline{q}_h = q_h^r + \alpha_h q_i^s$. Then there exists a constant C independent of h such that*

$$|\alpha_h|\big\|\nabla u_h^s\big\|_\infty \le C.$$

Proof We begin by noting, that $u^r \in W^{2,t}(\Omega)$ by definition of q_h^r. In particular, due to [15, Theorem 2]

$$\big\|\nabla u^r - \nabla u_h^r\big\|_\infty \le ch^{\beta}\big\|q_h^r\big\|_Q \le ch^{\beta}\|\overline{q}_h\|_Q \le ch^{\beta}.$$

Hence $\|\nabla u_h^r\|_\infty \le C$ independent of $h \in (0, 1]$.

This yields

$$\left\|\alpha_h \nabla u_h^s\right\|_\infty = \left\|\nabla \overline{u}_h - \nabla u_h^r\right\|_\infty \le 1 + C$$

and thus the assertion. □

In a next step, we need to show that $\|\nabla u_h^s\|_\infty$ blows up with a certain rate. This appears to be clear, unfortunately the author could not find a citable source. This is why we need the following lemma.

Lemma 5.2 *Let $\omega > \pi$ be the angle of the nonconvex corner of Ω. Further, let $u_h^s \in V_h$ be given by (5.4). Then for any $\varepsilon > 0$ there exists a constant c such that for $h > 0$ sufficiently small it holds*

$$\left\|\nabla u_h^s\right\|_\infty \ge c h^{-1+\pi/\omega+\varepsilon}.$$

Proof Denote the nonconvex corner by v. Let $p > 1$ be given. Then it is well known, that the solution $u^s \in V$ of (5.3) satisfies

$$\max_{\operatorname{dist}(v,x)=h^{1/p}} u^s(x) \ge c_1 h^{\frac{\pi}{p\omega}}$$

for some given constant $c_1 > 0$ and h sufficiently small.

By [16, Theorem 4.1] we now that for any $\varepsilon' > 0$ there exists some $c_2 > 0$ such that

$$\left\|u^s - u_h^s\right\|_\infty \le c_2 h^{\frac{\pi}{\omega}-\varepsilon'}.$$

Then for given $p > 1$ it holds

$$c_2 h^{\frac{p-1}{p}\frac{\pi}{\omega}-\varepsilon'} < c_1/2$$

provided that h is sufficiently small.

In particular it holds for any $x \in \Omega$

$$u_h^s(x) \ge u^s(x) - c_2 h^{\frac{\pi}{\omega}-\varepsilon'}.$$

Hence we have for h sufficiently small that

$$\begin{aligned}\max_{\operatorname{dist}(v,x)=h^{1/p}} u_h^s(x) &\ge c_1 h^{\frac{\pi}{p\omega}} - c_2 h^{\frac{\pi}{\omega}-\varepsilon'}\\ &= h^{\frac{\pi}{p\omega}}\left(c_1 - c_2 h^{\frac{p-1}{p}\frac{\pi}{\omega}-\varepsilon'}\right)\\ &\ge \frac{c_1}{2} h^{\frac{\pi}{p\omega}}.\end{aligned}$$

On the other hand $u_h^s(v) = 0$. Thus we have

$$\begin{aligned}\max_{\Omega} \left|\nabla u_h^s(x)\right| &\geq \frac{\max_{\operatorname{dist}(v,x)=h^{1/p}} u_h^s(x) - u_h^s(v)}{h^{1/p}} \\ &\geq \frac{c_1}{2} h^{\frac{\pi}{p\omega}} h^{\frac{-1}{p}} \\ &= \frac{c_1}{2} h^{\frac{1}{p}(-1+\frac{\pi}{\omega})}.\end{aligned}$$

Now, for given $\varepsilon > 0$ such that $-1 + \frac{\pi}{\omega} + \varepsilon < 0$ there exists some $p > 1$ such that

$$-1 + \frac{\pi}{\omega} + \varepsilon = \frac{1}{p}\left(-1 + \frac{\pi}{\omega}\right).$$

This proofs the assertion. □

Corollary 5.1 *For any $\varepsilon > 0$ there exists a constant c such that for $h > 0$ sufficiently small the singular coefficients α_h satisfy*

$$|\alpha_h| \leq ch^{1-\pi/\omega-\varepsilon}$$

Proof The assertion follows immediately from Lemma 5.2 and Lemma 5.1. □

Lemma 5.3 *Let $(\overline{q}_h, \overline{u}_h)$ be the unique solution to (2.8a)–(2.8c). Define $u^h \in V$ as the solution to*

$$\left(\nabla u^h, \nabla\varphi\right) = (\overline{q}_h, \varphi) \quad \forall\varphi \in V.$$

Then it holds

$$\left\|\left(\left|\nabla u^h\right| - 1\right)^+\right\|^2 \leq ch^{2\pi/\omega}.$$

Proof By definition and the fact, that $|\nabla\overline{u}_h| \leq 1$, we have for almost all $x \in \Omega$

$$\begin{aligned}\left(\left|\nabla u^h(x)\right| - 1\right)^+ &= \max\left(0, \left|\nabla u^h(x)\right| - 1\right) \\ &\leq \max\left(0, \left|\nabla u^h(x) - \nabla\overline{u}_h\right| + |\nabla\overline{u}_h| - 1\right) \\ &\leq \max\left(0, \left|\nabla u^h(x) - \nabla\overline{u}_h\right|\right) \\ &= \left|\nabla u^h(x) - \nabla\overline{u}_h\right|.\end{aligned}$$

Hence we get by standard finite error estimates

$$\left\|\left(\left|\nabla u^h(x)\right| - 1\right)^+\right\|^2 \leq \left\|\nabla u^h - \nabla\overline{u}_h\right\|^2 \leq ch^{2\pi/\omega}.$$

□

Theorem 5.2 *Let* $(\overline{q}_h, \overline{u}_h)$ *be the unique solution to* (2.8a)–(2.8c) *and* $(\overline{q}, \overline{u})$ *be the unique solution to* (2.2a)–(2.2c)*. Then for h sufficiently small, there exists a constant* $c > 0$ *such that*

$$\left|J(\overline{q}, \overline{u}) - J(\overline{q}_h, \overline{u}_h)\right| \le ch^{\beta_2}$$

where

$$\beta_2 = \frac{2\beta}{4+3\beta}(1 - \pi/\omega - \varepsilon)$$

for any $\varepsilon > 0$.

Proof In view of Theorem 3.1 we already know that

$$J(\overline{q}_h, \overline{u}_h) \le J(\overline{q}, \overline{u}) + ch^{\beta}.$$

Hence it remains to derive a lower bound on $J(\overline{q}_h, \overline{u}_h)$. To this end, we define $u^h \in V$ by

$$\left(\nabla u^h, \nabla\varphi\right) = (\overline{q}_h, \varphi) \quad \forall\varphi \in V.$$

Now, by standard L^2-error estimates we have

$$\left\|u - u^h\right\| \le ch^{2\pi/\omega}$$

and because $2\pi/\omega \ge 1 > \beta$ we get

$$\begin{aligned} J\left(\overline{q}_h, u^h\right) &\le J(\overline{q}, \overline{u}) + ch^{\beta}, \\ \left|J\left(\overline{q}_h, u^h\right) - J(\overline{q}_h, \overline{u}_h)\right| &\le ch^{\beta}. \end{aligned} \tag{5.5}$$

From this we immediately see, that if

$$J(\overline{q}, \overline{u}) \le J\left(\overline{q}_h, u^h\right)$$

we would be done.

Hence, we will now assume that

$$J\left(\overline{q}_h, u^h\right) \le J(\overline{q}, \overline{u}).$$

Then we proceed by considering a regularized version of (2.2a)–(2.2c) namely (4.2). Now we have that

$$\left|J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma) - J(\overline{q}, \overline{u})\right| \le c\gamma^{\frac{-\beta}{\beta+2}}$$

following the result of Theorem 4.1 and Lemma 4.3.

Further, with respect to Lemma 5.3 and Corollary 5.1 we have that

$$\left|J\left(\overline{q}_h, u^h\right) - J_\gamma\left(\overline{q}_h, u^h\right)\right| \le c\gamma h^{2\pi/\omega} + c\gamma^2 h^{2(1-\pi/\omega-\varepsilon)}$$

and thus

$$\begin{aligned} J(\overline{q},\overline{u}) - c\gamma^{\frac{-\beta}{2+\beta}} &\le J_\gamma(\overline{q}_\gamma,\overline{u}_\gamma) \\ &\le J_\gamma\left(\overline{q}_h,u^h\right) \le J\left(\overline{q}_h,u^h\right) + c\gamma h^{2\pi/\omega} + c\gamma^2 h^{2(1-\pi/\omega-\varepsilon)} \\ &\le J(\overline{q},\overline{u}) + c\gamma h^{2\pi/\omega} + c\gamma^2 h^{2(1-\pi/\omega-\varepsilon)} + ch^\beta. \end{aligned}$$

Now, in order to obtain the best possible rate of convergence, we choose $\gamma = h^{-x}$ where $x \ge 0$ solves

$$\max_{x\ge 0}\min\left(x\frac{\beta}{2+\beta}, 2\frac{\pi}{\omega} - x, 2(1-\pi/\omega-\varepsilon) - 2x\right) = f^*. \tag{5.6}$$

To do so, we note that $2\pi/\omega > 2 - 2\pi/\omega - \varepsilon$ since $\omega \le 2\pi$. Hence the minimizer is obtained when the two terms $x\frac{\beta}{2+\beta} =$ and $2(1-\frac{\pi}{\omega}-\varepsilon) - 2x$ are equilibrated. This happens at

$$\overline{x} = \left(1 - \frac{\pi}{\omega} - \varepsilon\right)\frac{4+2\beta}{4+3\beta}$$

with the value

$$f^* = \overline{x}\frac{\beta}{2+\beta} = \frac{2\beta}{4+3\beta}(1-\pi/\omega-\varepsilon) = \beta_2 < \beta.$$

Thus we obtain

$$J(\overline{q},\overline{u}) - ch^{\beta_2} \le J\left(\overline{q}_h,u^h\right) \le J(\overline{q},\overline{u}) + ch^{\beta_2}$$

which shows the assertion. □

Convergence of the primal variables follows analog to Corollary 4.2 using the uniform convexity of J_γ to get that

$$\|\overline{q}_\gamma - \overline{q}_h\|_Q^r + \|\overline{u}_\gamma - \overline{u}_h\|^2 \le ch^{\beta_2}.$$

References

1. H. Blum, M. Dobrowolski, On finite element methods for elliptic equations on domains with corners. Computing **28**, 53–63 (1982)
2. S. Brenner, L.R. Scott, *The Mathematical Theory of Finite Element Methods*, 3rd edn. (Springer, New York, 2008)
3. E. Casas, J.F. Bonnans, Contrôle de systèmes elliptiques semilinéares comportant des contraintes sur l'état, in *Nonlinear Partial Differential Equations and their Applications 8*, ed. by H. Brezzis, J.L. Lions (Longman, New York, 1988), pp. 69–86
4. E. Casas, L.A. Fernández, Optimal control of semilinear elliptic equations with pointwise constraints on the gradient of the state. Appl. Math. Optim. **27**, 35–56 (1993)

5. K. Deckelnick, A. Günther, M. Hinze, Finite element approximation of elliptic control problems with constraints on the gradient. Numer. Math. **111**, 335–350 (2008)
6. J. Douglas Jr., T. Dupont, L. Wahlbin, The stability in L^q of the L^2-projection into finite element function spaces. Numer. Math. **23**, 193–197 (1974/75)
7. P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, 1st edn. Monographs and Studies in Mathematics (Pitman, Boston, 1985)
8. A. Günther, M. Hinze, Elliptic control problems with gradient constraints—variational discrete versus piecewise constant controls. Comput. Optim. Appl. **49**(3), 549–566 (2009)
9. M. Hintermüller, M. Hinze, Moreau-Yosida regularization in state constrained elliptic control problems: error estimates and parameter adjustment. SIAM J. Numer. Anal. **47**(3), 1666–1683 (2009)
10. M. Hintermüller, M. Hinze, R.H.W. Hoppe, Weak-duality based adaptive finite element methods for PDE-constrained optimization with pointwise gradient state-constraints. Preprint 2010-08, Hamburger Beiträge zur Angewandten Mathematik, 2010
11. M. Hintermüller, K. Kunisch, PDE-constrained optimization subject to pointwise constraints on the control, the state, and its derivative. SIAM J. Optim. **20**(3), 1133–1156 (2009)
12. M. Hintermüller, A. Schiela, W. Wollner, The length of the primal-dual path in Moreau-Yosida-based path-following for state constrained optimal control. Preprint 2012-03, Hamburger Beiträge zur Angewandten Mathematik, 2012
13. M. Hinze, A variational discretization concept in control constrained optimization: the linear-quadratic case. Comp. Optim. Appl. **30**(1), 45–61 (2005)
14. C. Ortner, W. Wollner, A priori error estimates for optimal control problems with pointwise constraints on the gradient of the state. Numer. Math. **118**(3), 587–600 (2011)
15. A.H. Schatz, A weak discrete maximum principle and stability of the finite element method in L_∞ on plane polygonal domains. I. Math. Comp. **33**(148), 77–91 (1980)
16. A.H. Schatz, L.B. Wahlbin, Maximum norm estimates in the finite element method on plane polygonal domains. Part 1. Math. Comp. **32**(141), 73–109 (1978)
17. A. Schiela, W. Wollner, Barrier methods for optimal control problems with convex nonlinear gradient state constraints. SIAM J. Optim. **21**(1), 269–286 (2011)
18. M. Ulbrich, *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems*. MOS-SIAM Series on Optimization (SIAM, Philadelphia, 2011)
19. W. Wollner, A posteriori error estimates for a finite element discretization of interior point methods for an elliptic optimization problem with state constraints. Comput. Optim. Appl. **47**(1), 133–159 (2010)
20. W. Wollner, Optimal control of elliptic equations with pointwise constraints on the gradient of the state in nonsmooth polygonal domains. SIAM J. Control Optim. **50**(4), 2117–2129 (2012)