Springer Optimization and Its Applications 109

Jean-Baptiste Hiriart-Urruty
Adam Korytowski
Helmut Maurer
Maciej Szymkat *Editors*

Advances in Mathematical Modeling, Optimization and Optimal Control

Springer

Springer Optimization and Its Applications

VOLUME 109

Aims and Scope

Optimization has been expanding in all directions at an astonishing rate during the last few decades. New algorithmic and theoretical techniques have been developed, the diffusion into other disciplines has proceeded at a rapid pace, and our knowledge of all aspects of the field has grown even more profound. At the same time, one of the most striking trends in optimization is the constantly increasing emphasis on the interdisciplinary nature of the field. Optimization has been a basic tool in all areas of applied mathematics, engineering, medicine, economics, and other sciences.

The series *Springer Optimization and Its Applications* publishes undergraduate and graduate textbooks, monographs and state-of-the-art expository work that focus on algorithms for solving optimization problems and also study applications involving such problems. Some of the topics covered include nonlinear optimization (convex and nonconvex), network flow problems, stochastic optimization, optimal control, discrete optimization, multi-objective programming, description of software packages, approximation techniques and heuristic approaches.

More information about this series at http://www.springer.com/series/7393

Jean-Baptiste Hiriart-Urruty • Adam Korytowski
Helmut Maurer • Maciej Szymkat
Editors

Advances in Mathematical Modeling, Optimization and Optimal Control

Springer

Editors
Jean-Baptiste Hiriart-Urruty
Institut de Mathématiques
Université Paul Sabatier
Toulouse, France

Helmut Maurer
Institute of Computational and Applied Mathematics
University of Münster
Münster, Germany

Adam Korytowski
Department of Automatics and Biomedical Engineering
AGH University of Science and Technology
Kraków, Poland

Maciej Szymkat
Department of Automatics and Biomedical Engineering
AGH University of Science and Technology
Kraków, Poland

ISSN 1931-6828 ISSN 1931-6836 (electronic)
Springer Optimization and Its Applications
ISBN 978-3-319-30784-8 ISBN 978-3-319-30785-5 (eBook)
DOI 10.1007/978-3-319-30785-5

Library of Congress Control Number: 2016939392

Mathematics Subject Classification: 49J15, 49J20, 49J21, 49J30, 49N90, 26B25, 49M29, 97N20, 47J35

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Contents

Introduction

Jean-Baptiste Hiriart-Urruty, Adam Korytowski, Helmut Maurer, and Maciej Szymkat

This book constitutes a collection of developed versions of plenary papers presented (with one exception) at the 16th French–German–Polish Conference on Optimization, held in Kraków in 2013. They are authored by researchers of international repute in the field of optimization and optimal control. The book includes a number of new theoretical results and applications in biomechanics, medical technology, image processing, robot control, etc.

The purpose of the book was to give the authors an opportunity to present their new results to a wider audience than it was possible at the conference, and in an extended, more comprehensive form. The motivation was that the topics of the articles are related to areas of theory and applications that are of most vivid interest to the scientific community, such as image processing, partial differential inclusions, shape optimization, optimal control in medical and rehabilitation technology, or sufficient conditions of optimality.

The first paper, by Martin Burger, provides an overview of recent developments related to Bregman distances. Approaches in inverse problems and image processing based on Bregman distances are discussed, which have evolved to a standard tool

J.-B. Hiriart-Urruty
Institut de Mathématiques, Université Paul Sabatier, 31062 Toulouse Cedex 09, France

A. Korytowski (✉)
Department of Automatics and Biomedical Engineering, AGH University of Science and Technology, 30-059 Kraków, Poland
e-mail: akor@agh.edu.pl

H. Maurer
University of Münster, Institute of Computational and Applied Mathematics, 48149 Münster, Germany

M. Szymkat
AGH University of Science and Technology, Department of Applied Computer Science, 30-059 Kraków, Poland

J.-B. Hiriart-Urruty et al. (eds.), *Advances in Mathematical Modeling, Optimization and Optimal Control*, Springer Optimization and Its Applications 109,
DOI 10.1007/978-3-319-30785-5_1

in these fields in the last decade. Related issues in the analysis of nonlinear partial differential equations with a variational structure are also considered.

The paper by Piotr Kalita studies the operator version of a first order in time partial differential inclusion and its time discretization by implicit Euler scheme. The semidiscrete trajectories are proved to converge to the solution of the original problem. It is shown that, as times goes to infinity, all trajectories are attracted towards the so-called global attractors. It is also proved that the semidiscrete attractors converge upper-semicontinuously to the global attractor of the time continuous problem.

In the paper by Günter Leugering, Jan Sokołowski, and Antoni Żochowski, nonsmooth shape optimization problems for variational inequalities are considered. The variational inequalities model elliptic boundary value problems with the Signorini type unilateral boundary conditions. The shape functionals are given by the first order shape derivatives of the elastic energy. The topological optimization is used for passive control of singularities of weak solutions. The Hadamard directional differentiability is employed to sensitivity analysis. The topological derivatives of nonsmooth integral shape functionals for variational inequalities are derived. The obtained expressions for derivatives prove useful in numerical optimization for contact problems.

The next paper, by Katja Mombaur, is devoted to applications of optimal control and inverse optimal control in the field of medical and rehabilitation technology, in particular in human movement analysis, therapy and improvement by means of medical devices. Efficient methods for the solution of optimal control and inverse optimal control problems are discussed. Example applications of these methods are considered in the development of mobility aids for geriatric patients, the design of exoskeletons, the analysis of running motions with prostheses, the optimal functional electrical stimulation of hemiplegic patients, as well as stability analysis.

The last paper, by Nikolai Osmolovskii and Helmut Maurer, provides a survey on no-gap second-order optimality conditions in the calculus of variations and optimal control, and a discussion of their further development. Such conditions are formulated for discontinuous controls in optimal control problems with endpoint and mixed state-control constraints, and a free control time. For problems with the control appearing linearly in the Pontryagin function, it is shown that the second-order sufficient condition for the Induced Optimization Problem together with the so-called strict bang-bang property ensure second-order sufficient conditions for the original control problem. The theoretical results are illustrated by three applications: to optimal control of chemotherapy of HIV, time-optimal control of robots, and control of the Rayleigh equation.

Bregman Distances in Inverse Problems and Partial Differential Equations

Martin Burger

Abstract The aim of this paper is to provide an overview of recent development related to Bregman distances outside its native areas of optimization and statistics. We discuss approaches in inverse problems and image processing based on Bregman distances, which have evolved to a standard tool in these fields in the last decade. Moreover, we discuss related issues in the analysis and numerical analysis of nonlinear partial differential equations with a variational structure. For such problems Bregman distances appear to be of similar importance, but are currently used only in a quite hidden fashion. We try to work out explicitly the aspects related to Bregman distances, which also lead to novel mathematical questions and may also stimulate further research in these areas.

1 Introduction

Bregman distances for (differentiable) convex functionals, originally introduced in the study of proximal algorithms in [11] and named in [25], are a well-established concept in continuous and discrete optimization in finite dimension. A classical example is the celebrated Bregman projection algorithm for finding points in the intersection of affine subspaces (cf., e.g., [26]). We refer to [26, 53] for introductory and exhaustive views on Bregman distances in optimization.

Although convex functionals play a role in many other branches of mathematics, e.g., in many variational problems and partial differential equations, the suitability of Bregman distances in such fields was hardly investigated for several decades. In mathematical imaging and inverse problems the situation changed with the rediscovery and further development of Bregman iterations as an iterative image restoration technique in the case of frequently used regularization techniques such as total variation (cf. [50]), which led to significantly improved results compared to standard variational models and could eliminate systematic errors to a certain extent

M. Burger (✉)
Institut für Numerische und Angewandte Mathematik, Westfälische Wilhelms-Universität (WWU) Münster. Einsteinstr. 62, D 48149 Münster, Germany
e-mail: martin.burger@wwu.de

J.-B. Hiriart-Urruty et al. (eds.), *Advances in Mathematical Modeling, Optimization and Optimal Control*, Springer Optimization and Its Applications 109,
DOI 10.1007/978-3-319-30785-5_2

(cf. [9, 16]). Another key observation increasing the interest in Bregman distances in these fields was that they can be employed for error estimation, in particular for not strictly convex and nonsmooth functionals (cf. [14]), which prevent norm estimates.

Although there are many obvious links to the main route of research in Bregman distances and related optimization algorithms, there are several peculiar aspects that deserve particular discussion. Besides missing smoothness of the considered functionals and the fact that problems in imaging, inverse problems and partial differential equations are naturally formulated in infinite-dimensional Banach spaces such as the space of functions of bounded variation or Sobolev spaces, which have only been considered in few instances before, a key point is that the motivation for using Bregman distances in these fields often differs significantly from those in optimization and statistics. In the following we want to provide an overview of such questions and consequent developments, keeping an eye on potential directions and questions for future research. We start with a section including definitions, examples, and some general properties of Bregman distances, before we survey aspects of Bregman distances in inverse problems and imaging developed in the last decade. Then we proceed to a discussion of Bregman distances in partial differential equations, which is less explicit and hence the main goal is to highlight hidden use of Bregman distances and make the idea more directly accessible for future research. Finally we conclude with a section on related recent developments.

2 Bregman Distances and Their Basic Properties

We start with a definition of a Bregman distance. In the remainder of this paper, let X be a Banach space and $J : X \to \mathbb{R} \cup \{+\infty\}$ be convex functionals. We first recall the definition of subdifferential, respectively, subgradients.

Definition 1. The subdifferential of a convex functional J is defined by

$$\partial J(u) = \{p \in X^* \mid J(u) + \langle p, v - u \rangle \leq J(v) \text{ for all } v \in X\}. \tag{1}$$

An element $p \in \partial J(u)$ is called subgradient.

Having defined a subdifferential we can proceed to the definition of Bregman distances, respectively, generalized Bregman distances according to [44].

Definition 2. The (generalized) Bregman distance related to a convex functional J with subgradient p is defined by

$$D_J^p(v,u) = J(v) - J(u) - \langle p, v - u \rangle, \tag{2}$$

where $p \in \partial J(u)$. The symmetric Bregman distance is defined by

$$D_J^{p,q}(u,v) = D_J^p(v,u) + D_J^q(u,v) = \langle p - q, u - v \rangle, \tag{3}$$

where $p \in \partial J(u)$, $q \in \partial J(v)$.

Note that in the differentiable case, i.e., $\partial J(u)$ being a singleton, we can omit the special subgradient and write $D_J(v,u)$ or $D_J^{J'(u)}(v,u)$.

By the definition of subgradients the nonnegativity is apparent:

Proposition 1. *Let J be convex and $p \in \partial J(u)$. Then*

$$D_J^p(v,u) \geq 0 \qquad \forall\, v \in X$$

and

$$D_J^p(u,u) = 0.$$

If J is strictly convex, then $D_J^p(v,u) > 0$ for $v \neq u$.

We can further characterize vanishing Bregman distances as sharing a subgradient:

Proposition 2. *Let J be convex and $p \in \partial J(u)$. Then $D_J^p(v,u) = 0$ if and only if $p \in \partial J(v)$.*

Since Bregman distances are convex with respect to the first argument, we can also compute a subdifferential with respect to that variable, which is simply a shift of the subdifferential of J:

Proposition 3. *Let J be convex, $p \in \partial J(u)$. Then*

$$\partial_v D_J^p(v,u) = \partial J(v) - p.$$

Concerning existence proofs for variational problems involving Bregman distance it is often useful to investigate lower semicontinuity properties. Since Bregman distances can be considered as affinely linear perturbations of the functional J it is natural that these properties carry over:

Proposition 4. *Let J be convex and $q \in \partial J(v)$. Then the functional H defined by*

$$H(u) = D_J^q(u,v)$$

is convex. Hence, if X is reflexive, then H is weakly lower semicontinuous. If X is the dual of some Banach space Z and J is the convex conjugate of a functional on Z, then $q \in Z$ implies that H is lower semicontinuous in the weak star topology.

2.1 Examples of Bregman Distances

In the following we provide several examples of Bregman distances as frequently found in literature as well as some that received recent attention. This shall provide further insights into the relation to other distance measures and the basic properties of Bregman distances:

Example 1. Let X be a Hilbert space and $J(u) = \frac{1}{2}\|u\|_X^2$. Then $\partial J(u) = \{u\}$ and hence

$$D_J^u(v,u) = \frac{1}{2}\|u-v\|_X^2. \tag{4}$$

Example 2. Let I be a countable index set and $X = \ell^1(I)$ with

$$J(u) = \|u\|_{\ell^1} = \sum_{i\in I} |u_i|.$$

Then the Bregman distance is given by

$$D_J^p(v,u) = \sum_{i\in I}(q_i - p_i)v_i = \sum_{i,v_i>0}(1-p_i)|v_i| + \sum_{i,v_i<0}(1+p_i)|v_i|. \tag{5}$$

Note that the above sums have nonzero entries only if the sign of u_i does not match the sign of v_i, since $p_i = 1$ if $u_i > 0$ and $p_i = -1$ if $u_i < 0$.

Example 3. Let $X = \ell^1_+(\{1,\dots,N\})$ with

$$J(u) = \sum_{i=1}^N u_i \log u_i + 1 - u_i,$$

which is called the logarithmic entropy (or Boltzmann entropy). Then the Bregman distance is given by

$$D_J^p(v,u) = \sum_{i=1}^N v_i \log \frac{v_i}{u_i} + u_i - v_i, \tag{6}$$

which is known as Kullback–Leibler divergence. An analogous treatment applies to $X = L^1_+(\Omega)$, for a bounded domain Ω, and the continuous version

$$J(u) = \int_\Omega \left(u(x)\log u(x) + 1 - u(x)\right)\, dx,$$

resulting in the Bregman distance

$$D_J^p(v,u) = \int_\Omega \left(v(x) \log \frac{v(x)}{u(x)} + u(x) - v(x) \right) dx. \tag{7}$$

2.2 Bregman Distances and Duality

Duality is a basic ingredient in convex optimization (cf. [32]) and hence it is also interesting to understand some connections of duality and Bregman distances. For this sake we employ the convex conjugate (also called Legendre–Fenchel transform) of a functional J given by $J^* : X^* \to \mathbb{R} \cup \{+\infty\}$ satisfying

$$J^*(p) = \sup_{u \in X} \left(\langle p, u \rangle - J(u) \right). \tag{8}$$

Noticing that for $p \in \partial J(u)$ we have $J^*(p) = \langle p, u \rangle - J(u)$ one can immediately rewrite the Bregman distance as

$$D_J^p(v,u) = J(v) + J^*(p) - \langle p, v \rangle, \tag{9}$$

which can be interpreted as measuring the deviation of p from being a subgradient in $\partial J(v)$ or the deviation of v from being a subgradient in $\partial J^*(p)$.

A key identity relates Bregman distances with respect to J to those with respect to the convex conjugate J^*:

Proposition 5. *Let $p \in \partial J(u)$ and $q \in \partial J(v)$. Then*

$$D_J^p(v,u) = D_{J^*}^v(p,q). \tag{10}$$

Proof. By simple reordering we find

$$\begin{aligned} D_J^p(v,u) &= J(v) - \langle p, v \rangle + \langle p, u \rangle - J(u) \\ &= J(v) - \langle p, v \rangle + J^*(p), \end{aligned}$$

where we have used the maximality relation for the convex conjugate, which is equivalent to $p \in \partial J(u)$. With analogous reasoning we find $J^*(q) = \langle q, v \rangle - J(v)$ and hence

$$D_J^p(v,u) = J(v) + J^*(p) - J^*(q) - \langle p - q, v \rangle = D_{J^*}^v(p,q),$$

noticing that $q \in \partial J(v)$ implies $v \in \partial J^*(q)$.

A second aspect of duality related to Bregman distance is the convex conjugate of the latter, which shows that Bregman distances are dual to measuring differences via a functional:

Proposition 6. *Let $q \in \partial J(v)$ and H be defined by*

$$H(u) = D_J^q(u,v). \tag{11}$$

Then

$$H^*(p) = J^*(p+q) - J^*(q). \tag{12}$$

Proof. We have

$$\begin{aligned} H^*(p) &= \sup_u \left[\langle p,u\rangle - J(u) + J(v) - \langle q, v-u\rangle\right] \\ &= \sup_u \left[\langle p+q,u\rangle - J(u)\right] - \left[\langle q,v\rangle - J(v)\right]. \end{aligned}$$

The first term equals $J^*(p+q)$ by definition and the second equals $J^*(q)$ since $q \in \partial J(v)$.

2.3 Bregman Distances and Fenchel Duality

In the following we further investigate some properties of Bregman distances for a combination of two convex functionals $F : X \to \mathbb{R} \cup \{+\infty\}$, $G : Y \to \mathbb{R} \cup \{+\infty\}$. The classical setting is related the Fenchel duality theorem (cf. [32]), where

$$J(u) := F(u) + G(Ku) \tag{13}$$

with $K : X \to Y$ a bounded linear operator between Banach spaces. The Fenchel duality theorem shows that under suitable conditions

$$\inf_u J(u) = \sup_w \left[F^*(-K^*w) + G^*(w)\right], \tag{14}$$

together with equations relating optimal solutions $\hat{u}$ and $\hat{w}$ via subdifferentials of the involved functionals

$$-K^*\hat{w} \in \partial F(\hat{u}), \qquad K\hat{u} \in \partial G^*(w). \tag{15}$$

The above duality opens the possibility to employ Bregman distances on the dual problem as well as on the primal, which is nicely complemented by the duality relations for Bregman distances of a functional and its convex conjugate.

In the following we derive a basic estimates for the variational problem (13), which clarifies the relation of perturbations of one functional with duality and Bregman distances. We shall assume that the regularity of F and G is such that

$$\partial J(u) = \partial F(u) + K^* \partial G(Ku)$$

and the Fenchel duality theorem holds (cf. [32] for details).

Then we obtain the following estimate for perturbations of J:

Theorem 7. *Let F, G and K be as above, and let $\tilde{G}$ be a perturbation of G satisfying the same assumptions. Let $u \in X$ be a minimizer of J with $-K^*w \in \partial F(u)$ and $\tilde{u}$ be a minimizer of $F(\cdot) + \tilde{G}(K\cdot)$ with $-K^*\tilde{w} \in \partial F(\tilde{u})$. Then*

$$D_F^{-K^*w,-K^*\tilde{w}}(u,\tilde{u}) \leq G^*(\tilde{w}) - G^*(w) + \tilde{G}^*(w) - \tilde{G}^*(\tilde{w}). \tag{16}$$

Proof. We have

$$\begin{aligned} D_F^{-K^*w,-K^*\tilde{w}}(u,\tilde{u}) &= \langle K^*\tilde{w} - K^*w, u - \tilde{u}\rangle \\ &= \langle Ku, \tilde{w} - w\rangle + \langle K\tilde{u}, w - \tilde{w}\rangle. \end{aligned}$$

By the Fenchel duality theorem we have $Ku \in \partial G^*(w)$ and $K\tilde{u} \in \partial G^*(\tilde{w})$, which implies the assertion by inserting the subgradient inequality.

2.4 Bregman Distances for One-Homogeneous Functionals

The case of convex one-homogeneous functionals J, i.e.,

$$J(tu) = |t|J(u) \qquad \forall\, t \in \mathbb{R}, \tag{17}$$

received strong attention recently, and also appears to be a particularly interesting one with respect to Bregman distances. In the one-homogeneous case one has

$$J(u) = \langle p, u\rangle \tag{18}$$

for $p \in \partial J(u)$. Thus, the Bregman distance simply reduces to

$$D_J^p(v,u) = J(v) - \langle p, v\rangle. \tag{19}$$

An interesting property in the one-homogeneous case is the fact that the convex conjugate is the indicator function of a convex set C, i.e.,

$$J^*(p) = \begin{cases} 0 & \text{if } p \in C \\ +\infty & \text{else.} \end{cases} \tag{20}$$

This sheds interesting light on (10), noticing that $p \in \partial J(u)$ implies $p \in C$. Hence,

$$D_J^p(v,u) = D_{J^*}^v(p,q) = \langle q-p, v\rangle.$$

An alternative way to see this property is (19) combined with $\langle q, v\rangle = J(v)$.

In the one-homogeneous case we immediately find an example of Bregman distances vanishing for $v \neq u$. Let $t > 0$ and $v = tu$, then $\partial J(v) = \partial J(u)$ implies $D_J^p(v,u) = 0$. On the other hand, we observe that the Bregman distance distinguishes different orientation. Choosing $v = tu$ for $t < 0$ we have $\partial J(v) = -\partial J(u)$, hence $D_J^p(v,u) = 2J(v)$.

3 Applications in Inverse Problems and Imaging

In the last decade, Bregman distances have become an important tool in inverse problems and image processing. Their main use is twofold: On the one hand, they are of particular importance for all kinds of error estimates as already sketched above and in particular they are quite useful for the analysis of variational regularization techniques with non-differentiable regularization functionals. This route has been initiated in [14] and subsequently expanded, e.g., in [8, 18, 34, 36–38, 52, 54, 58]. On the other hand, Bregman distances can be used to construct novel iterative techniques with superior properties compared to classical variational regularization. This route goes back to [50] and was developed further, e.g., in [12, 16, 22, 35, 47, 60, 61, 63], the methods also had a huge impact on various applications (cf., e.g., [12, 28, 49]).

The basic setup we consider is the solution of a problem of the form $Ku = f$, where $K : X \to Y$ is a bounded linear operator between Banach spaces and f are given data. Since in most cases K does not have a closed range (or is even a compact operator) and data contain measurement errors, this problem can be ill-posed. To cure this issue variational regularization methods employ a convex regularization functional $R : X \to \mathbb{R} \cup \{+\infty\}$, which introduces the a-priori knowledge that reasonable approximations of the solution u have small (minimal) values $R(u)$. Variational regularization methods make a compromise between approximating the data f and minimizing R and solve a problem of the form

$$D(Ku,f) + \alpha R(u) \to \min_{u\in X}, \tag{21}$$

where $D : Y \times Y \to \mathbb{R}$ is an appropriate distance measure and $\alpha > 0$ is a regularization parameter to be chosen appropriately in dependence of the measurement error (often referred to as data noise). Specific forms of the distance measure are derived, e.g., via statistical modelling as the negative log-likelihood of the data noise. Frequently D is simply a least-squares term, i.e., Y is a Hilbert space and

$$D(Ku,f) = \frac{1}{2}\|Ku-f\|_Y^2. \tag{22}$$

A classical example is the ROF-model for image denoising [55], where R is the total variation seminorm, K is an embedding from $BV(\Omega)\cap L^2(\Omega)$ into $L^2(\Omega)$, and D the squared L^2-norm. For the whole section we shall assume that D is convex with respect to the first variable, which is the case for almost all commonly used examples.

3.1 Error Estimates

Error estimates for solutions of (21) are of interest with respect to two quantities: First of all, the distance of the data f to the ideal data Ku^*, where u^* is the unknown ideal solution. This part is referred to as data error or noise. Second, the regularization parameter α, which should equal zero in the case of ideal data and introduces a systematic error in the case of perturbed data (when it needs to be positive). In the setting of (13) we thus need to choose

$$F(u) = \alpha R(u), \qquad G(Ku) = D(Ku,f). \tag{23}$$

The optimality conditions for a minimizer u_α are then of the form

$$p_\alpha = K^* w_\alpha, \qquad p_\alpha \in \partial R(u_\alpha) \qquad -\alpha K^* w_\alpha \in \partial D(Ku_\alpha,f), \tag{24}$$

where the subgradient of D is meant to be computed with respect to the first argument for fixed f.

In order to obtain error estimates for some different data $\tilde f$ we choose $\tilde G(Ku) = D(Ku,\tilde f)$ and denote by $\tilde u_\alpha$ its corresponding regularized solution with

$$\tilde p_\alpha = K^* \tilde w_\alpha, \qquad \tilde p_\alpha \in \partial R(\tilde u_\alpha).$$

Then (16) yields

$$\alpha D_R^{K^* w_\alpha, K^* \tilde w_\alpha}(u,\tilde u) \le G^*(\tilde w_\alpha) - G^*(w_\alpha) + \tilde G^*(w_\alpha) - \tilde G^*(\tilde w_\alpha). \tag{25}$$

To further illustrate the behaviour consider the case of a quadratic data fidelity

$$G(Ku) = D(Ku,f) = \frac{1}{2}\|Ku - f\|^2, \tag{26}$$

for some squared Hilbert space norm, which yields $G^*(w) = \frac{1}{2}\|w\|^2 + \langle w, f\rangle$. Hence,

$$\alpha D_R^{K^* w_\alpha, K^* \tilde w_\alpha}(u,\tilde u) \le \langle f - \tilde f, \tilde w_\alpha - w_\alpha\rangle. \tag{27}$$

In the case (26) one can see quite immediately why the (symmetric) Bregman distance is an appropriate error measure for the estimates. Starting with the optimality conditions

$$Ku_\alpha - f + \alpha w_\alpha = 0, \qquad p_\alpha = K^* w_\alpha \in R(u_\alpha),$$
$$K\tilde{u}_\alpha - f + \alpha \tilde{w}_\alpha = 0, \qquad \tilde{p}_\alpha = K^* \tilde{w}_\alpha \in R(\tilde{u}_\alpha),$$

we find

$$K(u_\alpha - \tilde{u}_\alpha) + \alpha(w_\alpha - \tilde{w}_\alpha) = f - f^*. \tag{28}$$

The right-hand side is exactly the perturbation of the data, whose norm we want to use to estimate errors in the solution u_α. Hence we simply take the squared norm on both sides and obtain by expanding on the left-hand side

$$\|K(u_\alpha - \tilde{u}_\alpha)\|^2 + 2\alpha\langle w_\alpha - \tilde{w}_\alpha, K(u_\alpha - \tilde{u}_\alpha)\rangle + \alpha^2\|w_\alpha - \tilde{w}_\alpha\|^2 = \|f - \tilde{f}\|^2.$$

Finally using $K^* w_\alpha = p_\alpha$ we arrive at

$$\|K(u_\alpha - \tilde{u}_\alpha)\|^2 + 2\alpha D_R^{p_\alpha, \tilde{p}_\alpha}(u_\alpha, \tilde{u}_\alpha) + \alpha^2\|w_\alpha - \tilde{w}_\alpha\|^2 = \|f - \tilde{f}\|^2, \tag{29}$$

which implies (by the nonnegativity of all involved terms) the immediate estimate

$$D_R^{p_\alpha, \tilde{p}_\alpha}(u_\alpha, \tilde{u}_\alpha) \le \frac{1}{2\alpha}\|f - \tilde{f}\|^2 \tag{30}$$

for the Bregman distance. Note that (29) is not just an estimate, but indeed an equality for three error terms—the error in the image of the operator K (somehow the residual), the error in the dual variables w, and the Bregman distance of solutions. Here Ku and w are elements of a Hilbert space and it is of course natural to measure their deviations in the corresponding norm, so (29) yields the Bregman distance as the naturally induced error measure in the Banach space X.

Having obtained (29) it is interesting to note that one can alternatively obtain estimates for two parts of the right-hand side by taking scalar products of (28) with appropriate elements and subsequent application of the Cauchy–Schwarz, respectively, Young's inequality. The first is obtained by a scalar product with $Ku_\alpha - Ku^*$, which yields

$$\|K(u_\alpha - \tilde{u}_\alpha)\|^2 + \alpha D_R^{p_\alpha, \tilde{p}_\alpha}(u_\alpha, \tilde{u}_\alpha) = \langle f - \tilde{f}, K(u_\alpha - \tilde{u}_\alpha)\rangle \le \frac{1}{2}\|f - \tilde{f}\|^2 + \frac{1}{2}\|K(u_\alpha - \tilde{u}_\alpha)\|^2,$$

hence

$$\|K(u_\alpha - \tilde{u}_\alpha)\|^2 + 2\alpha D_R^{p_\alpha, \tilde{p}_\alpha}(u_\alpha, \tilde{u}_\alpha) \le \|f - \tilde{f}\|^2. \tag{31}$$

Using analogous reasoning, a scalar product of (28) with $w_\alpha - \tilde{w}_\alpha$ leads to

$$2\alpha D_R^{p_\alpha, \tilde{p}_\alpha}(u_\alpha, \tilde{u}_\alpha) + \alpha^2\|w_\alpha - \tilde{w}_\alpha\|^2 \le \|f - \tilde{f}\|^2. \tag{32}$$

3.2 Asymptotics

A key question in inverse problems is the behaviour of the regularized solution u_α as $\alpha \to 0$, which only makes sense if the noise in the data vanishes, i.e., $f = Ku^*$ for some desired solution $u^* \in X$. It is well-known that for ill-posed problems the convergence can be arbitrarily slow as $\alpha \to 0$ without further conditions on the desired solution u^*. For a further characterization it is important to note that under appropriate choice of α a limiting solution u^* of the variational model (21) satisfies

$$R(u) \to \min_{u \in X} \quad \text{subject to } Ku = Ku^*. \tag{33}$$

This can be seen from the estimate

$$D(Ku_\alpha, f) + \alpha R(u_\alpha) \le D(Ku^*, f) + \alpha R(u^*).$$

Using $\alpha \to 0$ and $D(Ku^*, f) \to 0$ we see that $D(Ku_\alpha, f) \to 0$, hence the limit is a solution of $Ku^* = f$. Dividing by α and using nonnegativity of D, we find

$$R(u_\alpha) \le R(u^*) + \frac{D(Ku^*, f)}{\alpha},$$

and under the standard condition on the parameter choice

$$\frac{D(Ku^*, f)}{\alpha} \to 0,$$

we observe that the limit of u_α cannot have a larger value of R than any other solution of $Ku = f$, i.e., it solves (33).

The key observation in [14, 27] is that appropriate conditions in the case of variational regularization is related to the existence of a Lagrange multiplier for (33). The Lagrange functional is given by $L(u, w) = R(u) - \langle w, Ku - Ku^* \rangle$, hence the existence of a Lagrange multiplier is the so-called *source condition*

$$p^* = K^* w^* \in \partial R(u^*). \tag{34}$$

Let us again detail the arguments in the case (26), where we can indeed use the above error estimates like (26) with $\tilde{u}_\alpha = u^*$ and $\tilde{w}_\alpha = w^*$. In order to obtain u_α as the solution of a variational problem we can indeed choose $\tilde{f} = Ku^* + \alpha w^*$ (note that (34) is equivalent to the existence of some $\tilde{f}$ such that u^* solves the variational problem with data $\tilde{f}$, cf. [14]). Hence, (29) becomes

$$\|K(u_\alpha - u^*)\|^2 + 2\alpha D_R^{p_\alpha, p^*}(u_\alpha, u^*) + \alpha^2 \|w_\alpha - w^*\|^2 = \|f - Ku^* - \alpha w^*\|^2. \tag{35}$$

Again with Young's inequality we end up at

$$D_R^{p_\alpha,p^*}(u_\alpha,u^*) \leq \frac{\|f-Ku^*\|^2}{\alpha} + \alpha\|w^*\|^2, \tag{36}$$

which gives the usual optimal choice $\alpha \sim \|f-Ku^*\|$ of regularization parameter in terms of the noise level, exactly as in the linear Hilbert space case (cf. [33]).

3.3 Bregman Iterations and Inverse Scale Space Methods

A frequent observation made for variational methods as discussed above is a systematic bias, in particular the methods yield solutions u_α with $R(u_\alpha)$ too small, which, e.g., results into a local loss of contrast in the case of total variation regularization (the contrast loss is larger for smaller structures cf. [17]). In order to cure such systematic errors in particular in the case of one-homogeneous regularization it turned out that the well-known Bregman iteration is a perfect tool. Instead of solving the variational problem only once one usually starts at u_0 being a minimizer of the regularization functional R, i.e., at the coarsest scale (if one agrees that scale is defined by R). Then of course $p_0 = 0 \in \partial R(u_0)$ and one can subsequently iterate

$$u_{k+1} \in \arg\min_{u\in X} \left(D(Ku,f) + \alpha D_R^{p_k}(u,u_k)\right), \tag{37}$$

where the subgradient p_k is updated via the optimality condition

$$p_{k+1} - p_k \in -\frac{1}{\alpha}K^*\partial D(Ku_k,f). \tag{38}$$

Noticing that we can again write $p_k = K^*w_k$ one can also construct an iteration

$$w_{k+1} - w_k \in -\frac{1}{\alpha}\partial D(Ku_k,f), \tag{39}$$

from which one can derive the well-known equivalence to augmented Lagrangian methods for minimizing R subject to $Ku = f$.

The convergence analysis in the case $f = Ku^*$ follows the well-known route for the Bregman iteration, but due to the ill-posedness of $Ku = f$ there is a particularly interesting aspect in the case of noisy data f differing from the ideal Ku^*. If the range of K is not closed, one has to take care of the situation where neither a solution $Ku = f$ nor some kind of least-squares solution (a minimizer of $D(Ku,f)$) exists in X. Hence, the Bregman iteration has the role of an iterative regularization method and needs to be stopped appropriately before the noise effects start to deteriorate the quality of the solution. Indeed one can show that the Bregman distance $D^{p_k}(u^*,u_k)$ is decreasing during the first iterations up to a certain point when the residual $D(Ku_k,f)$ becomes too small (i.e. one approximates the noisy data stronger than

Ku^*). Successful stopping criteria as the discrepancy principle are indeed based on comparing the residual with an estimate of the noise $D(Ku^*,f)$ and stop when $D(u_k,f)$ drops below this estimate.

In imaging a particularly interesting and quite related aspect of Bregman iterations is the scale behaviour. As mentioned above, with scale defined as above by properties of the regularization functional R, the Bregman iteration inserts finer and finer scales during its progress. In order not to miss certain scales it is obviously interesting to make small enough steps, which amounts to choosing α sufficiently large. For the limit of $\alpha \to \infty$ one can interpret the iteration as a backward Euler discretization (with timestep $\frac{1}{\alpha}$) of a flow, which has been called inverse scale space method by a reminiscence to so-called scale space methods in image processing, which exhibit the opposite scale behaviour (cf. [57, 59]). The inverse scale space flow is a solution of the differential inclusion

$$\partial_t p(t) \in -K^* \partial D(Ku(t),f), \qquad p(t) \in \partial J(u(t)), \tag{40}$$

with initial value $u(0) = u_0$ such that $p(0) = 0 \in \partial R(u_0)$. It can be interpreted by a gradient flow for the subgradient p on a dual functional (cf. [17]) or as a doubly nonlinear evolution equation. For the latter we will give an explanation on the analysis in terms of Bregman distances related to the involved functionals in the next section, which is also the appropriate way to analyse the inverse scale space method.

An unexpected result is the behaviour of the inverse scale space flow for polyhedral functions such as the ℓ^1-norm. Roughly speaking the polyhedral case means that for any $u \in X$ a subdifferential $\partial R(u)$ can be obtained via convex combinations of a finite number of elements (independent of u). It has been shown (cf. [20, 48]) that in such cases and $D(Ku,f) = \frac{1}{2}\|Ku-f\|^2$ the dynamics of the solution $u(t)$ is piecewise constant in time, i.e., quite far from a continuous flow, while the dynamics of the subgradients $p(t)$ is piecewise linear in time. Interestingly, the time steps t_k at which the solution changes can be computed explicitly, and the value of $u(t_k)$ is obtained by minimizing

$$\|Ku-f\|^2 \quad \text{subject to } \mathrm{p}(t_k) \in \partial R(u).$$

This is particularly attractive in the case of sparse optimization with R being the ℓ^1-norm, since the condition $p(t_k) \in \partial R(u)$ defines the sign of u and in particular the set of zeros. This means that the least-squares problems have to be solved on a rather small support, which is highly attractive for computational purposes (cf. [20]). Let us briefly explain the behaviour for $R : \mathbb{R}^N \to \mathbb{R}^+$ being the ℓ^1-norm and some arbitrary differentiable functional G on the right-hand side, i.e.,

$$\partial_t p_i(t) = -\partial_{u_i} G(u(t)). \tag{41}$$

In this case the subdifferential is the multivalued sign of $u_i(t))$ and for $u_0 = p_0 = 0$ we obviously find $u_i(t) = 0$ for sufficiently small time since $|p_i(t)| < 1$, which holds for all i. Hence for $t < t_1$ with t_1 to be determined we find

$$\partial_t p_i(t) = -\partial_{u_i} G(0), \tag{42}$$

which can be integrated easily to

$$p_i(t_1) = -t_1 \partial_{u_i} G(0). \tag{43}$$

The key observation is that $u_i \neq 0$ for some i is only possible if $|p_i(t_1)| = 1$. This implies that the first time with possibly nonzero u is

$$t_1 = \frac{1}{\|\partial G(0)\|_\infty}. \tag{44}$$

At time t_1 the sign of all u_i is determined by $p_i(t_1)$ and one can check that a solution is obtained by minimizing

$$u(t_1) \in \arg\min_{u \in \mathbb{R}^N} G(u) \quad \text{subject to } p_i(t_1) \in \partial |u_i(t_1)|, \tag{45}$$

or in other words

$$u(t_1) \in \arg\min_{u \in \mathbb{R}^N} G(u) \quad \text{subject to } p_i(t_1) u_i(t_1) \geq |u_i(t_1)| \ \forall i.$$

The optimality condition for the latter problem can be written as

$$\partial_{u_i} G(u(t_1)) + \lambda_i (q_i - p_i(t_1)) = 0, \qquad q_i \in \partial |u_i(t_1)|. \tag{46}$$

for some $\lambda \in \mathbb{R}^N$ satisfying the complementarity conditions

$$\lambda_i \geq 0, \qquad \lambda_i (p_i(t_1) u_i(t_1) - |u_i(t_1)|) = 0.$$

This implies $\partial_{u_i} G(u(t_1)) = 0$ of $u_i(t_1) \neq 0$, $\partial_{u_i} G(u(t_1)) \geq 0$ if $u_i(t_1) = 0$ and $p_i(t_1) = 1$, and $\partial_{u_i} G(u(t_1)) \leq 0$ if $u_i(t_1) = 0$ and $p_i(t_1) = -1$. This implies that we can find a time interval (t_1, t_2) such that

$$u(t) = u(t_1), \qquad p(t) = p(t_1) - (t - t_1) \partial G(u(t_1))$$

is a solution, and t_2 is again defined as the minimal time where there exists i such that $|p_i(t_2)| = 1$ and $|p_i(t)| < 1$. Again, the solution at time t_2 is defined by a solution of the variational problem

$$u(t_2) \in \arg\min_{u \in \mathbb{R}^N} G(u) \quad \text{subject to } p_i(t_2) u_i(t_2) \geq |u_i(t_2)| \ \forall i.$$

By an inductive procedure one obtains that the same kind of dynamics goes on for all t until it stops after finite time steps t_n at a minimizer of G.

As mentioned above the scale behaviour of the inverse scale space flow is highly attractive in image processing. In the polyhedral case there is a somehow exact decomposition into different scales by the steps made at times t_k. Indeed $\partial_t u$ is a sum of concentrated measures in time, and one may eliminate certain scales by leaving out the corresponding jump $u(t_k+\tau)-u(t_k-\tau)$. This observation leads the way to a much more general definition of filters from the inverse scale space method, which was discussed in [21]

$$\partial_t p(t) = f - u(t), \qquad p(t) \in \partial R(u(t)). \tag{47}$$

A certain scale filter is defined by

$$F(f) = u_0 + \int_0^\infty w(t) d\partial_t u(t), \tag{48}$$

with measurable weights $w(t) \in [0,1]$. In the case $w \equiv 1$ one simply obtains f, while certain scales can be damped out choosing $w(t) = 0$ for t in an appropriate interval. The design of filters for certain purpose is an ongoing subject of research.

4 Applications in Partial Differential Equations

In the following, we provide an overview of different aspects of partial differential equations, where Bregman distances are a useful tool. Unlike the case of inverse problems and image processing discussed above the notion of Bregman distance is not used widely in this field, and indeed most applications do not refer to this term or use it in a very hidden way. Our goal in the following section is to work out the basic ideas related to Bregman distances in a structured way, which sheds new light on many established techniques and hopefully also opens routes towards novel results. For this sake we employ a formal approach and avoid technicalities such as detailed function spaces, which of course can be worked out from existing literature.

4.1 Entropy Dissipation Methods for Gradient Systems

Entropy dissipation methods are a frequently used tool in partial differential equations (cf. [4, 41]), which is often based on using the logarithmic entropy

$$E(u) = \int_\Omega u(x) \log u(x)\, dx \tag{49}$$

as a Lyapunov functional, e.g., in diffusion equations (cf., e.g., [2–5, 24]), kinetic equations (cf., e.g., [4]), or fluid mechanics (cf., e.g., [56]). In particular in gradient

systems also different convex functionals are used regularly and in a structured way. The abstract form of a gradient system is

$$\partial_t u(t) = -L(u(t))E'(u(t)), \tag{50}$$

where $L(u)$ is a linear symmetric positive semi-definite operator on appropriate spaces and E a convex energy functional, which we assume differentiable for simplicity (similar treatment for non-differentiable convex functionals is possible by using subgradients, but beyond our scope). The entropy dissipation property can be verified by the straightforward computation

$$\frac{d}{dt}E(u(t)) = E'(u(t))\partial_t u(t) = -\langle E'(u(t)), L(u(t))E'(u(t))\rangle \leq 0. \tag{51}$$

The negative of the right-hand side is frequently called entropy dissipation functional $D(u(t))$ and can be used to derive further quantitative information about the decay to equilibrium. A standard example (cf. [4, 24]) are nonlinear Fokker–Planck equations of the form

$$\partial_t u = \nabla \cdot (m(u)\nabla(e'(u) + V)) \tag{52}$$

on a domain $\Omega \subset \mathbb{R}^d$ with no-flux boundary conditions. Here, $e : \mathbb{R}^+ \to \mathbb{R}$ is a convex function, $m : \mathbb{R}^+ \to \mathbb{R}^+$ a (potentially nonlinear) mobility function, and $V : \Omega \to \mathbb{R}$ an external potential. Recently also systems of Fokker–Planck equations as well as certain reaction–diffusion systems of the form

$$\partial_t u_i = D_i \Delta u_i + F_i(u_1, \ldots, u_M), \qquad i = 1, \ldots, M \tag{53}$$

have been investigated with entropy dissipation techniques (cf. [29, 42, 45]).

The major purpose of entropy dissipation techniques is to obtain qualitative or ideally quantitative results about the decay to equilibrium for transient solutions. An equilibrium solution u_∞ is a minimizer of E on a convex set K, to which also the transient solution $u(t)$ belongs for all t. An example is the Fokker–Planck equation with linear mobility $m(u) = u$, where K is the set of nonnegative integrable functions with prescribed mean value. Hence, u_∞ satisfies

$$E'(u_\infty)(u - u_\infty) \geq 0 \qquad \forall\, u \in K. \tag{54}$$

If further the operator $L(u)$ is such that

$$L(u)E'(u_\infty) = 0 \qquad \forall\, u \in K, \tag{55}$$

which is indeed the case for the typical examples, then one can rewrite the gradient system as

$$\partial_t u(t) = -L(u(t))(E'(u(t)) - E'(u_\infty)). \tag{56}$$

Hence, the right-hand side is expressed in a difference of energy gradients for the transient and equilibrium solution. In a similar way, the entropy dissipation can be rewritten in terms of a distance between those and the Bregman distance (usually called relative entropy) plays a key role for this purpose. One observes that

$$\begin{aligned}\frac{d}{dt} D_E^{E'(u_\infty)}(u(t), u_\infty) &= E'(u(t))\partial_t u(t) - E'(u_\infty)\partial_t u(t) \\ &= -\langle E'(u(t)) - E'(u_\infty), L(u(t))(E'(u(t)) - E'(u_\infty))\rangle \\ &=: -F(u(t), u_\infty).\end{aligned}$$

Of course, the above computation holds for smooth solutions only, for weak solutions one can usually derive the time-integrated version

$$D_E^{E'(u_\infty)}(u(t), u_\infty) + \int_s^t F(u(\tau)\, d\tau \leq D_E^{E'(u_\infty)}(u(s), u_\infty). \tag{57}$$

The above computation shows that entropy dissipation can be rephrased as the decrease of the Bregman distance between stationary and transient solution. We notice that the use of the Bregman distance is not essential in this computation, but the understanding of this structure can be quite beneficial, in particular if one wants to use dual variables, the so-called entropy variables

$$\varphi(t) = E'(u(t)), \qquad \varphi_\infty = E'(u_\infty). \tag{58}$$

The entropy variable φ solves the system

$$\partial_t (E'(\varphi(t)) = -L((E^*)'(\varphi(t)))\varphi(t), \tag{59}$$

where E^* is the convex conjugate of E. When analysing the dual flow (59) a dissipation property can now be derived immediately using relation (10). Thus, we obtain a dual entropy dissipation of the form

$$\frac{d}{dt} D_{E^*}^{u(t)}(\varphi_\infty, \varphi(t)) = --\langle \varphi(t) - \varphi_\infty, L((E^*)'(\varphi(t)))(\varphi(t) - \varphi_\infty)\rangle. \tag{60}$$

The duality relation is particularly interesting for constructing approximations in terms of the entropy variables, as, e.g., carried out for degenerate cross-diffusion systems in (cf. [19, 40, 62]).

4.2 Lyapunov Functionals for Gradient Systems Out of Equilibrium

The appropriate use of Bregman distances seems to be less explored, but maybe even more crucial for the derivation of Lyapunov functionals if gradient systems are perturbed out of equilibrium. The simplest example is the linear Fokker–Planck equation with non-potential force as investigated in [5]

$$\partial_t u = \nabla \cdot (\nabla u + uF) \qquad \text{in } \Omega \times \mathbb{R}^+, \tag{61}$$

supplemented by no-flux boundary conditions

$$(\nabla u + uF) \cdot n = 0 \qquad \text{on } \partial\Omega \times \mathbb{R}^+. \tag{62}$$

If the vector field F is not the gradient of some potential function, then a stationary solution cannot be constructed as the minimizer of an entropy functional. However, the existence and uniqueness of a stationary solution can be shown under quite general assumptions on F (cf. [31]). In a form similar to gradient flows (cf. [1, 53]) we write (61) as

$$\partial_t u = \nabla \cdot (u(\nabla e'(u) + F)), \qquad e(u) = u \log u + 1 - u, \tag{63}$$

which suggests to further investigate distances based on the entropy functional

$$E(u) = \int_\Omega e(u)\, dx = \int_\Omega (u \log u - u + 1)\, dx. \tag{64}$$

The dissipation of the relative entropy can be computed via

$$\begin{aligned}
\frac{d}{dt} D_E^{E'(u_\infty)}(u(t), u_\infty) &= \int_\Omega (e'(u(t)) - e'(u_\infty)) \partial_t u(t)\, dx \\
&= \int_\Omega (e'(u(t)) - e'(u_\infty)) \nabla \cdot u(\nabla(e'(u(t)) \\
&\quad - e'(u_\infty)) + \nabla e'(u_\infty) + F)\, dx \\
&= -\int_\Omega u |\nabla(e'(u(t)) - e'(u_\infty))|^2\, dx \\
&\quad + \int_\Omega (e'(u(t)) - e'(u_\infty)) \nabla \cdot u(t) (\nabla e'(u_\infty) + F)\, dx,
\end{aligned}$$

where we have used the no-flux boundary conditions

$$(\nabla e(u(t)) + F) \cdot n = (\nabla e(u_\infty) + F) \cdot n = 0 \qquad \text{on } \partial\Omega \times \mathbb{R}^+$$

in order to apply integration by parts in the first term on the right-hand side. The second term is simplified via

$$\begin{aligned}
&\nabla \cdot (u(\nabla e'(u_\infty) + F)) \\
&\quad = \nabla \cdot (\frac{u(t)}{u_\infty} u_\infty (\nabla e'(u_\infty) + F)) \\
&\quad = u_\infty \nabla (\frac{u(t)}{u_\infty}) \cdot (\nabla e'(u_\infty) + F)) \\
&\quad = u_\infty \nabla \exp(e'(u(t)) - e'(u_\infty)) \cdot (\nabla e'(u_\infty) + F)) \\
&\quad = u_\infty \exp(e'(u(t)) - e'(u_\infty)) \nabla (e'(u(t)) - e'(u_\infty)) \cdot (\nabla e'(u_\infty) + F)).
\end{aligned}$$

With Ψ satisfying $\Psi'(z) = z\exp(z)$ we can further write

$$\begin{aligned}
&\int_\Omega (e'(u(t)) - e'(u_\infty)) \nabla \cdot u(t) (\nabla e'(u_\infty) + F)\, dx = \\
&\qquad \int_\Omega \nabla \Psi (e'(u(t)) - e'(u_\infty)) \cdot u_\infty (\nabla e'(u_\infty) + F)\, dx = 0,
\end{aligned}$$

which can be seen again through integration by parts. Hence, we finally obtain the decrease of the Bregman distance via

$$\frac{d}{dt} D_E^{E'(u_\infty)}(u(t), u_\infty) = -\int_\Omega u |\nabla (e'(u(t)) - e'(u_\infty))|^2\, dx, \tag{65}$$

and the logarithmic Sobolev inequality (cf. [3]) implies exponential convergence to equilibrium.

Another example are boundary-driven nonlinear Fokker–Planck equations

$$\partial_t u = \nabla \cdot (\nabla m(u) + m(u) F) \qquad \text{in } \Omega \times \mathbb{R}^+, \tag{66}$$

considered in [10] with Dirichlet boundary conditions

$$u = g \qquad \text{on } \partial\Omega \times \mathbb{R}^+. \tag{67}$$

We mention that an analogous analysis holds in the case of no-flux boundary conditions (in which case we have a direct generalization of the nonsymmetric Fokker–Planck equation above) or mixed Dirichlet and no-flux boundary conditions. Bodineau et al. [10] construct Lyapunov functionals of the form

$$H(u, u_\infty) = \int_\Omega \int_{u_\infty(x)}^{u(x,t)} \Phi' \left(\frac{m(s)}{m(u_\infty(x))} \right) ds\, dx, \tag{68}$$

where Φ is a nonnegative function with unique minimum at zero. Such a construction seems far from being intuitive, but it becomes much more clear for Φ being the logarithmic entropy, i.e., $\Phi'(t) = \log t$. In this case the Lyapunov functional becomes

$$H(u, u_\infty) = \int_\Omega \int_{u_\infty(x)}^{u(x,t)} \log m(s) - \log m(u_\infty(x))\, ds\, dx, \tag{69}$$

and with a function e such that $e'(s) = \log m(s)$, we further obtain

$$H(u, u_\infty) = \int_\Omega (e(u(x,t)) - e(u_\infty(x)) - e'(u_\infty(x))(u(x,t) - u_\infty(x)))\, ds\, dx, \tag{70}$$

which is nothing but the Bregman distance for the entropy functional

$$E(u) = \int_\Omega e(u)\, dx, \qquad \text{with } e'(u) = \log m(u). \tag{71}$$

Since Eq. (66) can be written as

$$\partial_t u = \nabla \cdot (m(u)(\nabla \log m(u) + F)), \qquad \text{in } \Omega \times \mathbb{R}^+, \tag{72}$$

the above form of E is also a natural choice. The detailed computations for the entropy dissipation are indeed completely analogous to the case of the linear Fokker–Planck equation, the crucial point appears to be the logarithmic relation between entropy derivatives $e'(u)$ and mobilities $m(u)$.

4.3 Doubly Nonlinear Evolution Equations

A generalization of gradient systems are doubly nonlinear evolution equations with a gradient structure either of the form

$$\partial_t p(t) \in -\partial G(u(t)), \qquad p(t) \in \partial F(u(t)) \tag{73}$$

or as

$$\partial F(\partial_t u) + \partial G(u(t)) \ni 0. \tag{74}$$

The best studied case, which is also the one where both coincide, corresponds to $F(u) = \frac{1}{2}\|u\|^2$ for a norm in a Hilbert space, which yields the classical gradient flow

$$\partial_t u(t) \in -\partial G(u(t)). \tag{75}$$

We have seen a system in the form (73) already above in the inverse scale space method, while the form (74) appears frequently in mechanical problems (cf., e.g., [46] and the references therein). There is indeed a duality relation for (73) and (74). Starting from (73) we obtain $u(t) \in \partial G^*(-\partial_t p(t)) \cap \partial F^*(p(t))$, respectively $-u(t) \in \partial G^*(\partial_t p(t))$ if G satisfies a symmetry condition around zero. This yields

$$\partial G^*(\partial_t p(t)) + \partial F^*(p(t)) \neq 0,$$

the analogue of (74).

Doubly nonlinear evolution equations have recently been investigated extensively, and in particular tools from convex analysis have been employed (cf. [46]). Here we add our Bregman distance point of view to derive estimates for such equations. Let us start with a straightforward computation on the change of the time derivative of the Bregman distance:

Lemma 1. *Let F be differentiable and u a solution of (73). Then*

$$\frac{d}{dt} D_F^{p(t)}(v, u(t)) = -\langle \partial_t p(t), v - u(t) \rangle \leq G(v) - G(u(t)).$$

This can be used to quantify the distance of $u(t)$ to a minimizer of G:

Corollary 1. *Let F be differentiable, u_∞ a minimizer of G, and u a solution of (73). Then*

$$\frac{d}{dt} D_F^{p(t)}(u_\infty, u(t)) + D_G^0(u(t), u_\infty) \leq 0. \tag{76}$$

Since it is straightforward to see

$$\frac{d}{dt} D_G^0(u(t), u_\infty) = \frac{d}{dt} G(u(t)) \leq 0 \tag{77}$$

we see after integrating (76) in time

$$D_F^{p(t)}(u_\infty, u(t)) + t D_G^0(u(t), u_\infty) \leq D_F^{p(t)}(u_\infty, u(t)) + \int_0^t D_G^0(u(s), u_\infty)\, ds \leq D_F^{p(0)}(u_\infty, u(0)),$$

leading to linear decay of the Bregman distance:

Theorem 2. *Let F be differentiable, u_∞ a minimizer of G, and u a solution of (73). Then*

$$D_G^0(u(t), u_\infty) \leq \frac{1}{t} D_F^{p(0)}(u_\infty, u(0)). \tag{78}$$

4.4 Error Estimates for Nonlinear Elliptic Problems

We finally turn our attention to the analysis of discretization methods for nonlinear elliptic problems such as the p-Laplace equation. Such elliptic problems are optimality conditions of some energy functional of the form

$$E(u) = J(u) - \langle f, u \rangle, \tag{79}$$

where J is a convex functional on a Banach space X, typically a Sobolev space of first order derivatives. The elliptic differential equation (or more general differential inclusion) is the optimality condition

$$p = f, \qquad p \in \partial J(u) \tag{80}$$

A canonical example is the p-Laplace equation

$$-\nabla \cdot \left(|\nabla u|^{p-2} \nabla u \right) = f, \tag{81}$$

which is related to the functional

$$J(u) = \frac{1}{p} \int_\Omega |\nabla u(x)|^p \, dx. \tag{82}$$

For variational discretizations of such problems the Bregman distance appears to be a quite useful tool, which is still not fully exploited. In many approaches the Bregman distance is used in a hidden way and strict convexity is used to obtain an estimate in terms of the underlying norms (with potentially suboptimal constants, however). For the p-Laplace equation such an approach is carried out in [30]. Again in the limiting case $p = 1$ related to total variation minimization the Bregman distance is even more crucial and appears, e.g., in [6]. Here we briefly sketch the obvious role of Bregman distances in Galerkin discretizations of the form

$$E(u) \to \min_{u \in X_h}, \tag{83}$$

where X_h is a finite-dimensional subspace of X, e.g., constructed by finite elements.

Let us start by pointing out the basic structure of error estimates for Galerkin methods in the linear case related to the minimization of a positive definite quadratic form

$$J(u) = B(u, u), \tag{84}$$

where $B : X \times X \to \mathbb{R}$ is a bounded and coercive bilinear form. The optimality condition in weak form is given by

$$B(u,v) = \langle f, v\rangle \qquad \forall\, v \in X, \tag{85}$$

and the Galerkin discretization yields a solution $u_h \in X_h$ of

$$B(u_h,v) = \langle f, v\rangle \qquad \forall\, v \in X_h. \tag{86}$$

Error estimates for such discretizations are obtained in two steps: first the error between u and u_h is estimated by the projection error to the subspace X_h and then the projection error is estimated, e.g., via the interpolation error. The crucial property for the first step is the so-called *Galerkin orthogonality*

$$B(u-u_h,v) = 0 \qquad \forall\, v \in X_h, \tag{87}$$

which implies

$$B(u-u_h,u-u_h) = B(u-u_h,u-v) \qquad \forall\, v \in X_h, \tag{88}$$

and by the Cauchy–Schwarz inequality for the positive definite bilinear form B

$$B(u-u_h,u-u_h) \le B(u-v,u-v) \qquad \forall\, v \in X_h. \tag{89}$$

In other words u_h is the projection of u on the subspace X_h, when the (squared) norm induced by B is used as a distance measure.

Since the term $B(u-v,u-v)$ above is just the Bregman distance related to quadratic functional J one might think of an analogous property in the case of nonquadratic J, when the Bregman projection is used. Indeed, we can derive such a relation in the case of arbitrary convex J. For this sake let again u be a minimizer of E and u_h a minimizer of E constrained to the subspace X_h. Then we have $f \in \partial J(u)$ and thus, since u_h minimizes E on X_h, we have for all $v \in X_h$

$$\begin{aligned} D_J^f(u_h,u) &= J(u_h) - J(u) - \langle f, u_h - u\rangle \\ &= E(u_h) - J(u) + \langle f, u\rangle \\ &\le E(v) - J(u) + \langle f, u\rangle. \end{aligned}$$

Rewriting the last term we hence obtain the Bregman projection property

$$D_J^f(u_h,u) \le D_J^f(v,u), \qquad \forall\, v \in X_h. \tag{90}$$

This observation opens a way to analyse Galerkin methods for such nonlinear problems in the same way as in the linear case, the key step to be developed for specific problems and specific discretizations (X_h) is the estimation of the Bregman projection error.

Note again the role of the Bregman distance for error estimation: The one-sided distance $D_J^f(u_h,u)$ is particularly suitable for the estimation of a-priori errors as above, while a-posteriori error estimation should rather be based on the distance $D_J^{p_h}(u,u_h)$ with $p_h \in \partial J(u_h)$. We have by the minimizing property of u

$$\begin{aligned} D_J^{p_h}(u,u_h) &= J(u) - J(u_h) - \langle p_h, u - u_h \rangle \\ &= E(u) - E(u_h) + \langle p_h - f, u_h - u \rangle \\ &\leq \langle p_h - f, u_h - u \rangle. \end{aligned}$$

Using the duality relation $u \in \partial J^*(f)$, this could be further estimated to the full a-posteriori estimate

$$D_J^{p_h}(u,u_h) \leq \langle p_h - f, u_h \rangle + J^*(2f - p_h) - J^*(f). \tag{91}$$

For practical purposes the above abstract estimate is not useful in most cases, since computing the adjoint J^* means to solve a nonlinear partial differential equation as well, which might be as difficult as the original one. However, the general strategy can be exploited together with specific properties of the functional J and the subspace X_h. In particular for gradient energies of the form

$$J(u) = \int_\Omega j(\nabla u)\, dx \tag{92}$$

one can derive alternative versions using only the convex conjugate j^*, which is significantly easier to compute.

5 Further Developments

In this final section we discuss some aspects of Bregman distances that came up recently and will potentially have strong further impact, in particular we will explore some developments related to probability.

5.1 Uncertainty Quantification in Inverse Problems

Since Bregman distances appear to be a suitable tool for estimates in certain nonlinear deterministic problems, it seems natural to exploit them also in the stochastic counterparts of such problems. The obvious measure for error estimates is then the expected value of the Bregman distance with respect to the stochastic quantity. Such approaches have been used successfully in particular in statistical

inverse problems (cf., e.g., [58]), which we also want to discuss in the following. In order to avoid technicalities we restrict ourselves to a purely finite-dimensional setup.

Consider the inverse problem $Ku = f$, where $K : \mathbb{R}^N \to \mathbb{R}^M$ and the data are generated from a true solution u^* with additive Gaussian noise, i.e.,

$$f = Ku^* + \sigma n, \tag{93}$$

with n a Gaussian random variable with zero mean and covariance matrix I_M. Let again R be a convex regularization functional and u_α a solution of the variational problem

$$J(u) = \frac{1}{2\sigma^2}\|(Ku - f)\|^2 + \alpha J(u) \to \min_{u\in\mathbb{R}^N}. \tag{94}$$

Then u_α satisfies the optimality condition

$$\frac{1}{\sigma^2}K^*K(u_\alpha - u^*) + \alpha p_\alpha = \frac{1}{\sigma^2}K^*n, \qquad p_\alpha \in \partial R(u_\alpha), \tag{95}$$

which implies $p_\alpha = K^*w_\alpha$. Now assume u^* satisfies the source condition (34), then we have

$$K(u_\alpha - u^*) + \alpha\sigma^2(w_\alpha - w^*) = n - \alpha\sigma^2 w^*.$$

Taking the squared norm and subsequently expectation with respect to w in this identity we obtain

$$\begin{aligned} 2\alpha\sigma^2 E[D_R^{p_\alpha,p^*}(u_\alpha,u^*)] &\le E[\|K(u_\alpha - u^*)\|^2 + 2\alpha\sigma^2 D_R^{p_\alpha,p^*}(u_\alpha,u^*) + \alpha^2\sigma^4\|w_\alpha - w^*\|]^2 \\ &= E[\|n - \alpha\sigma^2 w^*\|^2] \\ &= E[\|n\|^2] + \alpha^2\sigma^4\|w^*\|^2 = \sigma^2 M + \alpha^2\sigma^4\|w^*\|^2. \end{aligned}$$

Thus, the expected error in the Bregman distance is estimated by

$$E[D_R^{p_\alpha,p^*}(u_\alpha,u^*)] \le \frac{M}{2\alpha} + \frac{\alpha\sigma^2}{2}\|w^*\|^2. \tag{96}$$

We notice that the above approach not only yields an estimate of the Bregman distance, but also indeed an exact value for the sum of three error measures, in addition to the Bregman distance also the residual error as well as the error in the source space (related to $w_\alpha - w^*$). Usually the latter is the largest of the three, so one needs to expect a blow up of this term as $M \to \infty$ if α is not increasing as M. If one is interested in the first two terms only, one can simply use a duality product with

$u_\alpha - u^*$ in (95) and subsequently estimate the expected value of the right-hand side in a different way, which may lead to robust estimates in terms of M, respectively, estimates that can be carried out for infinite-dimensional white noise.

An application of Bregman distances in Bayesian modelling was recently investigated in [13], considering frequently used posterior densities of the form

$$\pi(u|f) \sim e^{-\frac{\|Ku-f\|^2}{2\sigma^2} - \alpha R(u)}, \tag{97}$$

where again R is a convex and Lipschitz continuous functional on $\mathbb{R}^N$ (generalizations to posterior distributions in infinite-dimensional spaces where further studied in [39]). It has been shown that the posterior can be centred around the so-called maximum a-posteriori probability (MAP) estimate $\hat{u}$, which maximizes $p(u|f)$, in the form

$$\pi(u|f) \sim e^{-\frac{\|Ku-K\hat{u}\|^2}{2\sigma^2} - \alpha D_R^{\hat{p}}(u,\hat{u})}. \tag{98}$$

Based on the observation

$$\langle s, u-\hat{u}\rangle = \frac{\|Ku-K\hat{u}\|^2}{\sigma^2} + \alpha\langle p-\hat{p}, u-\hat{u}\rangle. \tag{99}$$

for $p \in \partial R(u)$ and

$$s = \frac{1}{\sigma^2}K^*(Ku-f) + \alpha p \in \partial(-\log \pi(u|f)),$$

a Bayes cost of the form

$$\Gamma(v) = \mathbf{E}_{p(u|f)}\left[\frac{\|Ku-Kv\|^2}{\sigma^2} + \alpha\langle q-\hat{p}, v-\hat{u}\rangle\right] \tag{100}$$

has been introduced for $q \in \partial R(v)$ (note that selection of $p \in R(u)$ is only needed on a set of zero measure due to Rademacher's theorem). A simple integration by parts argument then shows that the MAP-estimate $\hat{u}$ is a minimizer of the Bayes cost, which is a quite natural choice compared to the highly degenerate cost usually used to characterize MAP estimates (cf. [43]). A direct consequence is the fact that the MAP estimate has smaller Bregman distance in expectation than the frequently used conditional mean estimate, hence one obtains a theoretical argument explaining the success of MAP estimates in practice.

5.2 Bregman Distances and Optimal Transport

Bregman distances can be used also as a cost in optimal transport, which has been investigated in [23] for a convex and differentiable functional J on $\mathbb{R}^N$. Given two probability measures μ and ν, an optimal transport plan is a probability measure γ on $\mathbb{R}^N \times \mathbb{R}^N$ with marginals μ and ν minimizing the functional

$$F(\gamma) = \int_{\mathbb{R}^N \times R^N} D_J^{J'(u)}(v,u)\, d\gamma(v,u). \tag{101}$$

The resulting optimal value of F can be interpreted as a transport distance between the measures μ and ν.

Besides the important question of well-posedness solved in (cf. [23]) there are several interesting problems such as the existence of transport maps under certain condition (i.e. concentration of γ on a set described by the graph of a map $T : \mathbb{R}^N \to \mathbb{R}^N$) as well as relations to uncertainty quantification. A first example is the Bayes cost approach described in the previous section, which can indeed be interpreted as the transport distance between the posterior distribution and a measure concentrated at the MAP estimate. This motivates further research in the future, an obvious next step might be to estimate distances between different posterior distributions in transport distances related to Bregman distances.

A different use of Bregman distances in optimal transport was recently made in [7] for the solution of Monge–Kantorovich formulations in optimal transport. They consider entropic regularizations of the problem, i.e., for $\epsilon > 0$ they minimize a discrete version of

$$F_\epsilon(\gamma) = \int_{\mathbb{R}^N \times R^N} C(v,u)\, d\gamma(v,u) + \epsilon E(\gamma), \tag{102}$$

where E is the entropy

$$E(\gamma) = \int_{\mathbb{R}^N \times R^N} \log\left(\frac{d\gamma}{d\mathcal{L}}\right) d\gamma(v,u), \tag{103}$$

where $\frac{d\gamma}{d\mathcal{L}}$ is the Radon–Nikodym derivative with respect to the Lebesgue measure. The key observation is that the minimization of F_ϵ can be rewritten equivalently as the minimization of the Kullback–Leibler divergence, i.e., the Bregman distance related to E, between γ and the Gibbs measure φ_ϵ with density $e^{-C/\epsilon}$

$$D_E(\gamma, \varphi_\epsilon) \to \min_\gamma, \tag{104}$$

which transforms the problem into a Bregman projection problem of the Gibbs density onto the set of plans with given marginals, which can be computed much more efficiently than the original transport control problem. Note that the general procedure can be carried out as well with an arbitrary convex functional whose

domain are positive densities, the corresponding Gibbs density is then to be defined as $\varphi_\epsilon = (E^*)'(-C/\epsilon)$. A particular computational advantage of the logarithmic entropy is the fact that iterative Bregman projections can be computed explicitly and realized with low complexity, in the discrete sets it only needs multiplications and scalar products of diagonal matrices with the matrix discretizing the Gibbs measure (cf. [7] for further details).

5.3 Infimal Convolution of Bregman Distances

Infimal convolution of convex functionals become popular recently in image processing in order to combine favourable properties of certain regularization functionals, e.g., total variation and higher-order versions thereof. A quite unexplored topic is the infimal convolution of Bregman distances, however. Since they are convex functionals of the first variable one may consider the infimal convolution

$$[D_R^{p_1}(\cdot,u_1)\square D_R^{p_2}(\cdot,u_2)](u) = \inf_{v\in X}[D_R^{p_1}(u-v,u_1) + D_R^{p_2}(v,u_2)], \tag{105}$$

with an obvious extension to more than two values.

Of particular interest in imaging applications appears to be the case of $p_2 = -p_1$ and $u_2 = -u_1$ for a one-homogeneous functional such as total variation. The latter was used to obtain a regularization functional enforcing partly equal edge sets (REF colorbregman). While minimizing the Bregman distance for total variation strongly favours edge sets with jumps of equal sign (see also the discussion related to orientation for one-homogeneous functionals in Sect. 2.4), the infimal convolution of Bregman distances eliminates this part and hence measures differences in edge sets rather than jumps of the same sign. A further study of theoretical properties as well as applications of such kind of infimal convolution of Bregman distances remains an interesting property for future research. Obvious candidates are problems in compressed sensing where one first of all aims at obtaining the correct support of the solution rather than the sign.

Acknowledgements This work was partially supported by ERC via Grant EU FP 7 - ERC Consolidator Grant 615216 LifeInverse and by the German Science Foundation DFG via BU 2327/6-1 and EXC 1003 Cells in Motion Cluster of Excellence, Münster, Germany

References

1. Ambrosio, L., Gigli, N., Savare, G.: Gradient Flows in Metric Spaces and in the Space of Probability Measures. Birkhaeuser, Basel/Boston/Berlin (2005)
2. Arnold, A., Unterreiter, A.: Entropy decay of discretized Fokker-Planck equations I: temporal semidiscretization. Comput. Math. Appl. **46**, 1683–1690 (2003)

3. Arnold, A., Markowich, P., Toscani, G., Unterreiter, A.: On convex Sobolev inequalities and the rate of convergence to equilibrium for Fokker-Planck type equations. Commun. Partial Differ. Equ. **26**, 43–100 (2001)
4. Arnold, A., Carrillo, J.A., Desvillettes, L., Dolbeault, J., Jüngel, A., Lederman, C., Markowich, P., Toscani, G., Villani, C.: Entropies and equilibria of many-particle systems: an essay on recent research. Mon. Hefte Math. **142**, 35–43 (2004)
5. Arnold, A., Carlen, E., Ju, Q.: Large-time behavior of non-symmetric Fokker-Planck type equations. Commun. Stoch. Anal. **2**, 153–175 (2008)
6. Bartels, S.: Error control and adaptivity for a variational model problem defined on functions of bounded variation. Math. Comput. **84**, 1217–1240 (2015)
7. Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L.: Peyre, G., Iterative Bregman projections for regularized transportation problems. SIAM J. Sci. Comput. **37**, A1111–A1138 (2015)
8. Benning, M., Burger, M.: Error estimates for general fidelities. Electron. Trans. Numer. Anal. **38**, 44–68 (2011)
9. Benning, M., Burger, M.: Ground states and singular vectors of convex variational regularization methods. Methods Appl. Anal. **20**, 295–334 (2013)
10. Bodineau, T., Lebowitz, J., Mouhot, C., Villani, C.: Lyapunov functionals for boundary-driven nonlinear drift-diffusion equations. Nonlinearity **27**, 2111 (2014)
11. Bregman, L.M.: The relaxation method for finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Math. Phys. **7**, 200–217 (1967)
12. Brune, C., Sawatzky, A., Burger, M.: Primal and dual Bregman methods with application to optical nanoscopy. Int. J. Comput. Vis. **92**, 211–229 (2011)
13. Burger, M., Lucka, F.: Maximum-a-posteriori estimates in linear inverse problems with log-concave priors are proper Bayes estimators. Inverse Prob. **30**, 114004 (2014)
14. Burger, M., Osher, S.: Convergence rates of convex variational regularization. Inverse Prob. **20**, 1411–1421 (2004)
15. Burger, M., Osher, S.: A guide to the TV zoo. In: Burger, M., Osher, S. (eds.) Level Set and PDE-Based Reconstruction Methods in Imaging. Springer, Berlin (2013)
16. Burger, M., Gilboa, G., Osher, S., Xu, J.: Nonlinear inverse scale space methods. Commun. Math. Sci. **4**, 179–212 (2006)
17. Burger, M., Frick, K., Osher, S., Scherzer, O.: Inverse total variation flow. Multiscale Model. Simul. **6**, 366–395 (2007)
18. Burger, M., Resmerita, E., He, L.: Error estimation for Bregman iterations and inverse scale space methods in image restoration. Computing **81**, 109–135 (2007)
19. Burger, M., Di Francesco, M., Pietschmann, J.F., Schlake, B.: Nonlinear cross-diffusion with size exclusion. SIAM J. Math. Anal. **42**, 2842–2871 (2010)
20. Burger, M., Moeller, M., Benning, M., Osher, S.: An adaptive inverse scale space method for compressed sensing. Math. Comput. **82**, 269–299 (2013)
21. Burger, M., Eckardt, L., Gilboa, G., Moeller, M.: Spectral representation of one-homogeneous functionals. In: Aujol, J.F., Niholovq, M., Pzpadikis, N. (eds.) Scale Space and Variation Methods in Computer Vision, Springer, Berlin, Heibelly, 16–27
22. Cai, J., Osher, S., Shen, Z.: Linearized Bregman iterations for compressed sensing. Math. Comput. **78**, 1515–1536 (2009)
23. Carlier, G., Jimenez, C.: On Monge's problem for Bregman-like cost functions. J. Convex Anal. **14**, 647–655 (2007)
24. Carrillo, J.A., Jüngel, A., Markowich, P., Toscani, G., Unterreiter, A.: Entropy dissipation methods for degenerate parabolic problems and generalized Sobolev inequalities. Mon. Hefte Math. **133**, 1–82 (2001)
25. Censor, Y., Lent, A.: An iterative row-action method for interval convex programming J. Optim. Theory Appl. **34**, 321–353 (1981)
26. Censor, Y., Zenios, S.: Parallel Optimization: Theory, Algorithms, and Applications. Oxford University Press, Oxford (1998)

27. Chavent, G., Kunisch, K.: Regularization of linear least squares problems by total bounded variation. ESAIM: Control Optim. Calc. Var. **2**, 359–376 (1997)
28. Choi, K., Fahimian, B.P., Li, T., Suh, T.S., Lei, X.: Enhancement of four-dimensional cone-beam computed tomography by compressed sensing with Bregman iteration. J. X-Ray Sci. Technol. **21**, 177–192 (2013)
29. Di Francesco, M., Fellner, K., Markowich, P.: The entropy dissipation method for inhomogeneous reaction-diffusion systems. Proc. R. Soc. A **464**, 3272–3300 (2008)
30. Diening, L., Kreuzer, C.: Linear convergence of an adaptive finite element method for the p-Laplacian equation. SIAM J. Numer. Anal. **46**, 614–638 (2008)
31. Droniou, J., Vazquez, J.L.: Noncoercive convection-diffusion elliptic problems with Neumann boundary conditions. Calc. Var. Partial Differ. Equ. **34**, 413–434 (2009)
32. Ekeland, I., Temam, R.: Convex Analysis and Variational Problems. SIAM, Philadelphia (1999)
33. Engl, H., Hanke-Bourgeois, M., Neubauer, A.: Regularization of Inverse Problems. Kluwer, Dordrecht (1996)
34. Flemming, J.: Theory and examples of variational regularization with non-metric fitting functionals. J. Inverse III-Posed Probl. **18**, 677–699 (2010)
35. Goldstein, T., Osher, S.: The split Bregman method for l_1 regularized problems. SIAM J. Imaging Sci. **2**, 323–343 (2008)
36. Grasmair, M.: Generalized Bregman distances and convergence rates for non-convex regularization methods. Inverse Problems **11**, 115014 (2010)
37. Grasmair, M.: Variational inequalities and higher order convergence rates for Tikhonov regularisation on Banach spaces. J. Inverse III-Posed Probl. **21**, 379–394 (2013)
38. Hein, T.: Tikhonov regularization in Banach spaces - improved convergence rates results. Inverse Prob. **25**, 035002 (2009)
39. Helin, T., Burger, M.: Maximum a posteriori probability estimates in infinite-dimensional Bayesian inverse problems. Preprint arXiv:1412.5816 (2015)
40. Jüngel, A.: The boundedness-by-entropy method for cross-diffusion systems. Nonlinearity **28**, 1963–2001 (2015)
41. Jüngel, A., Matthes, D.: Entropiemethoden für nichtlineare partielle Differentialgleichungen. Int. Math. Nachr. **209**, 1–14 (2008)
42. Jüngel, A., Stelzer, I.V.: Entropy structure of a cross-diffusion tumor-growth model. Math. Mod. Meth. Appl. Sci. **22**, 1250009, (2012)
43. Kaipio, J., Somersalo, E.: Statistical and Computational Inverse Problems. Springer, New York (2005)
44. Kiwiel, K.C.: Proximal minimization methods with generalized Bregman functions. SIAM J. Control Optim. **35**, 1142–1168 (1997)
45. Mielke, A., Haskovec, J., Markowich, P.A.: On uniform decay of the entropy for reaction-diffusion systems. J. Dyn. Differ. Equ. **27**, 897–928 (2015)
46. Mielke, A., Rossi, R., Savare, G.: Nonsmooth analysis of doubly nonlinear evolution equations. Calc. Var. Partial Differ. Equ. **46**, 253–310 (2013)
47. Moeller, M.: Multiscale methods for polyhedral regularizations and applications in high dimensional imaging. Ph.D. thesis, University of Muenster (2012)
48. Moeller, M., Burger, M.: Multiscale methods for polyhedral regularizations. SIAM J. Optim. **23**, 1424–1456 (2013)
49. Müller, J., Brune, C., Sawatzky, A., Kösters, T., Schäfers, K.P., Burger, M.: Reconstruction of short time PET scans using Bregman iterations. In: Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), pp. 2383–2385. IEEE, New York (2011)
50. Osher, S., Burger, M., Goldfarb, D., Xu, J., Yin, W.: An iterative regularization method for total variation-based image restoration. SIAM Multiscale Model. Simul. **4**, 460–489 (2005)
51. Otto, F.: The geometry of dissipative evolution equations: the porous medium equation. Commun. Partial Differ. Equ. **26**, 101–174 (2001)
52. Pöschl, C.: An overview on convergence rates for Tikhonov regularization methods for non-linear operators. J. Inverse III-posed Probl. **17**, 77–83 (2009)

53. Reid, M.: Meet the Bregman divergences. http://mark.reid.name/blog/meet-the-bregman-divergences.html (2013)
54. Resmerita, E.: Regularization of ill-posed problems in Banach spaces: convergence rates. Inverse Problems **21**, 1303 (2005)
55. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Phys. D: Nonlinear Phenom. **60**, 259–268 (1992)
56. Saint-Raymond, L.: Convergence of solutions to the Boltzmann equation in the incompressible Euler limit. Arch. Ration. Mech. Anal. **166**, 47–80 (2003)
57. Scherzer, O., Groetsch, C.: Inverse scale space theory for inverse problems. In: Scale-Space and Morphology in Computer Vision, pp. 317–325. Springer, Berlin/Heidelberg (2001)
58. Werner, F., Hohage, T.: Convergence rates in expectation for Tikhonov-type regularization of inverse problems with Poisson data. Inverse Problems **28**, 104004 (2012)
59. Witkin, A.P.: Scale-space filtering: a new approach to multi-scale description. In: IEEE International Conference on ICASSP'84 Acoustics, Speech, and Signal Processing, vol. 9, pp. 150–153 (1984)
60. Xu J., Osher, S.: Iterative regularization and nonlinear inverse scale space applied to wavelet-based denoising. IEEE Trans. Image Process. **16**, 534–544 (2007)
61. Yin, W.: Analysis and generalizations of the linearized Bregman method. SIAM J. Imaging Sci. **3**, 856–877 (2010)
62. Zamponi, N., Jüngel, A.: Analysis of degenerate cross-diffusion population models with volume filling, TU Vienna. Preprint (2015)
63. Zhang, X., Burger, M., Bresson, X., Osher, S.: Bregmanized nonlocal regularization for deconvolution and sparse reconstruction. SIAM J. Imaging Sci. **3**, 253–276 (2010)

On Global Attractor for Parabolic Partial Differential Inclusion and Its Time Semidiscretization

Piotr Kalita

Abstract In this article we study the operator version of a first order in time partial differential inclusion as well as its time discretization obtained by an implicit Euler scheme. This technique, known as the Rothe method, yields the semidiscrete trajectories that are proved to converge to the solution of the original problem. While both the time continuous problem and its semidiscretization can have nonunique solutions we prove that, as times goes to infinity, all trajectories are attracted towards certain compact and invariant sets, so-called global attractors. We prove that the semidiscrete attractors converge upper-semicontinuously to the global attractor of time continuous problem.

1 Introduction

In the study of the long time behavior of solutions of initial and boundary value problems for dissipative partial differential equations or inclusions one can either study the asymptotic behavior of individual trajectories, or, like we do in this article, study the asymptotic behavior of sets reachable from the bounded sets of initial conditions.

This second approach leads to the theory of global attractors, the sets which are compact and invariant (or semiinvariant) in the phase space and attract all bounded sets of initial conditions. To know that such object, global attractor, exists, would mean that the dynamics is asymptotically captured by the solution map behavior on a certain compact set, while all trajectories outside this set become attracted towards it as time advances to infinity.

For problems governed by dissipative partial differential equations with uniqueness of solutions, theory of global attractors has been thoroughly studied (see [8, 9, 32, 36]). The problems without uniqueness of solution, in turn, have focused much attention recently (see [2]). There are two approaches to deal with the

P. Kalita (✉)
Faculty of Mathematics and Computer Science, Jagiellonian University, ul. prof. S. Łojasiewicza 6, 30-348 Kraków, Poland
e-mail: piotr.kalita@ii.uj.edu.pl

J.-B. Hiriart-Urruty et al. (eds.), *Advances in Mathematical Modeling, Optimization and Optimal Control*, Springer Optimization and Its Applications 109,
DOI 10.1007/978-3-319-30785-5_3

problems without uniqueness. One of them is based on the study of multifunctions that assign to the initial state the set of states reachable after some time t. This approach, by the so-called multivalued semiflows or m-semiflows was initiated in [1] and developed in [25, 26], or, more recently, in [11, 16, 40]. The second technique, by trajectory attractors, consists in the study of shift operators on the spaces of time dependent trajectories and was developed in [7, 24, 34].

The first approach was used in [21] to study the inclusion

$$u'(t) \in \Delta u(t) + F(u(t)),$$

where one of the assumptions on F is that it is Lipschitz continuous in the Hausdorff metric on $\mathbb{R}$, i.e.

$$\text{haus}_{\mathbb{R}}(F(u), F(v)) \leq L_F|u - v| \quad \text{for all} \quad u, v \in \mathbb{R}.$$

The authors prove that both the original problem and its semidiscretization (as well as fully time and space discretized scheme) generate m-semiflows that have global attractors, and the discrete attractors converge upper-semicontinuously to the attractor of the original, time continuous, problem.

In this article we obtain a similar result to that of [21] for a more general setup: namely, the multifunction does not have to be Hausdorff continuous, we require only strong–weak upper-semicontinuity. Moreover, this multifunction does not have to be defined on the state space of the problem, it can also originate from multivalued Neumann-Robin boundary condition, these types of boundary conditions are useful when modeling friction [23] or other contact phenomena [29]. We use the unified framework proposed in [15], where the multivalued term can be defined either on the boundary on inside the domain of interest. Finally, the operator in our problem does not have to be Laplacian, it must be pseudomonotone: a general class of not necessarily potential operators that include, for example, p-Laplacian or quasilinear operators in divergence form satisfying appropriate Leray–Lions conditions (see [33]).

Note that we prove here only upper-semicontinuous convergence of semidiscrete attractors to the time continuous one. This means that the attractors cannot "implode," i.e. all cluster points of semidiscrete attractors must belong to the time continuous one, however, it remains unknown, whether there are points in time continuous attractor, that are not the cluster points of semidiscrete ones, hence "explosion" of attractors is possible. We note here that the question of lower-semicontinuous convergence, i.e. whether such "explosion" is indeed possible is open and only some partial and unsatisfactory responses have been developed so far (see [2, 30] for the recent review).

The existence of global attractor for m-semiflow governed by the inclusion with pseudomonotone operator was proved in [17, 19, 20]. Moreover, a strong convergence of semidiscrete solutions to the time continuous one established in Theorem 5 below and needed for the convergence of attractors was used in [39]

Theorem 11.1 for non-monotone autonomous evolution inclusions in strongest topologies of the extended phase space (see also Theorems 4.1 and 4.2 from [14] for non-autonomous case and Sect. 6 of [18]).

The plan of the article is the following. In Sect. 2 some preliminary necessary definitions and results are recalled. Next Sect. 3 contains the definitions of the time continuous and time discretized problems as well as necessary assumptions. In the following Sect. 4 we prove the auxiliary result of pseudomonotonicity of Nemytskii operators. Section 5 is devoted to the convergence of the solutions of semidiscrete schemes, while Sects. 6 and 7 contain the proofs of the existence of time continuous and time discretized attractors, respectively. Convergence of discrete attractors to the time continuous one is proved in Sect. 8. The last Sect. 9 contains the discussion of the assumptions on the problem data and examples of operators that satisfy them.

2 Definitions

Definition 1. An operator $A : X \to X^*$, where X is a reflexive Banach space is pseudomonotone (in the sense of Brézis) if for every sequence $v_n \to v$ weakly in X such that $\limsup_{n\to\infty}\langle Av_n, v_n - v\rangle_{X^*\times X} \le 0$ we have

$$\langle Av, v - w\rangle_{X^*\times X} \le \liminf_{n\to\infty}\langle Av_n, v_n - w\rangle_{X^*\times X} \quad \text{for all} \quad w \in X.$$

Definition 2. Let X be a reflexive Banach space, and let $W \subset X$ be another Banach space embedded continuously in X. An operator $A : X \to X^*$ is W-pseudomonotone if for every sequence $\{v_n\} \subset W$ with $\|v_n\|_W$ bounded such that $v_n \to v$ weakly in X and $\limsup_{n\to\infty}\langle Av_n, v_n - v\rangle_{X^*\times X} \le 0$ we have

$$\langle Av, v - w\rangle_{X^*\times X} \le \liminf_{n\to\infty}\langle Av_n, v_n - w\rangle_{X^*\times X} \quad \text{for all} \quad w \in X.$$

The definition of pseudomonotonicity of multifunctions is not a simple generalization of single valued case.

Definition 3 (See [13], Chap. 1.3). A multifunction $A : X \to 2^{X^*}$, where X is a reflexive Banach space is multivalued pseudomonotone, if

(i) for any $v \in X$ the set $A(v)$ is nonempty, weakly compact, and convex,
(ii) A is upper-semicontinuous from every finite dimensional subspace of X into X^* furnished with weak topology,
(iii) if $v_n \to v$ weakly in X and $v_n^* \in A(v_n)$ satisfies $\limsup_{n\to\infty}\langle v_n^*, v_n - v\rangle \le 0$ then for every $y \in X$ there exists $u(y) \in A(v)$ such that

$$\langle u(y), v - y\rangle \le \liminf_{n\to\infty}\langle v_n^*, v_n - y\rangle.$$

Note that it is useful to check pseudomonotonicity of multifunctions via the following sufficient condition (see Proposition 1.3.66 in [13] or Proposition 3.1 in [6]).

Proposition 1. *Let X be a reflexive Banach space. A multifunction $A : X \to 2^{X^*}$ is multivalued pseudomonotone, if it satisfies the following conditions*

(i) for every $v \in X$ the set $A(v)$ is nonempty, weakly compact, and convex,
(ii) A is bounded,
(iii) if $v_n \to v$ weakly in X and $v_n^* \to v^*$ weakly in X^* with $v_n^* \in A(v_n)$ and if $\limsup_{n\to\infty} \langle v_n^*, v_n - v\rangle \leq 0$ then $v^* \in A(v)$ and $\langle v_n, v_n^*\rangle \to \langle v, v^*\rangle$.

Definition 4. Let $I = (a,b)$ be a finite time interval and let $u : I \to X$, where X is a Banach space. The q-variation seminorm is defined as

$$\|u\|^q_{BV^q(I;X)} = \sup\left\{ \sum_{i=0}^{k-1} \|u(t_{i+1}) - u(t_i)\|^q_X \mid k \in \mathbb{N}, a = t_0 < t_1 < t_2 < \ldots < t_k = b \right\}.$$

We denote by $BV^q(I;X)$ the set of all functions $u : I \to X$ for which the q-variation seminorm is finite.

For Banach spaces X, Y such that $X \subset Y$ we define the following Banach space, for $1 \leq p, q < \infty$ (see [15])

$$M^{p,q}(I;X,Y) = L^p(I;X) \cap BV^q(I;Y),$$

and if $u \in M^{p,q}(I;X,Y)$ for all bounded intervals $I \subset \mathbb{R}^+$ then we say that $u \in M^{p,q}_{loc}(\mathbb{R}^+;X,Y)$. We have the following theorem (see Theorem 1 in [15]) that motivates the use of space $M^{p,q}$.

Theorem 1. *Let $1 \leq p, q < \infty$. Let moreover $X_1 \subset X_2 \subset X_3$ be Banach spaces such that X_1 is reflexive, the embedding $X_1 \subset X_2$ is compact and the embedding $X_2 \subset X_3$ is continuous. If a set $\mathscr{S} \subset M^{p,q}(0,T;X_1,X_3)$ is bounded, then it is relatively compact in $L^p(0,T;X_2)$.*

We remind a definition of Painlevé-Kuratowski upper convergence of sets, and a useful theorem on pointwise convergence of weakly convergent sequences (see, for instance, Proposition 4.7.44 in [12])

Definition 5. Let (X,τ) be a topological space and let $\{A_n\}_{n=1}^{\infty}$ be a sequence of subsets of X. The upper Painlevé-Kuratowski limit of the sequence $\{A_n\}$ is defined by

$$\tau - \limsup_{n\to\infty} A_n = \{x \in X \,|\, x = \tau - \lim_{k\to\infty} x_{n_k}, x_{n_k} \in A_{n_k}, n_1 < n_2 < \ldots < n_k < \ldots\}.$$

Theorem 2. *Let I be an open and finite time interval and let X be a reflexive Banach space. Let $f \in L^1(I;X)$ and let $\{f_n\}_{n=1}^{\infty}$ be a sequence of functions from*

$L^1(I;X)$ such that $f_n \to f$ weakly in $L^1(I;X)$. Moreover, assume that for a.e. $t \in I$ we have $\|f_n(t)\|_X \leq R_t$, where R_t is a positive constant independent of n, but possibly dependent on $t \in I$. Then we have

$$f(t) \in \overline{\text{conv}} \text{ weak} - \limsup_{n\to\infty} \{f_n(t)\} \quad \text{for a.e.} \quad t \in I.$$

If $(X, \|\cdot\|_X)$ is a Banach space, then we define the distance from the point $x \in X$ to the set $A \subset X$ as $\text{dist}_X(x,A) = \inf_{y\in A} \|x-y\|_X$. Moreover we define the Hausdorff semidistance from the set $A \subset X$ to the set $B \subset X$ as $\text{dist}_X(A,B) = \sup_{x\in A} \text{dist}_X(x,B)$. Let $P(X)$ $(\mathscr{B}(X), C(X), K(X))$ be the family of all nonempty (nonempty and bounded, nonempty and closed, nonempty and compact) subsets of X. We denote $\mathbb{R}^+ = [0,\infty)$. Let $\mathbb{T}$ be the additive subgroup of $\mathbb{R}$ and $\mathbb{T}^+ = \mathbb{R}^+ \cap \mathbb{T}$.

Definition 6. The mapping $\mathscr{G} : \mathbb{T}^+ \times X \to P(X)$ is called a multivalued semiflow (m-semiflow) if

(*i*) $\mathscr{G}(0,x) = \{x\}$ for all $x \in X$,
(*ii*) $\mathscr{G}(t_1+t_2,x) \subset \mathscr{G}(t_1,\mathscr{G}(t_2,x))$ for all $x \in X$, $t_1, t_2 \in \mathbb{T}^+$, where for $A \subset X$ and $t \in \mathbb{T}^+$ we define $\mathscr{G}(t,A) = \bigcup_{x\in A} \mathscr{G}(t,x)$.

If in (*ii*), instead of inclusion, the equality $\mathscr{G}(t_1+t_2,x) = \mathscr{G}(t_1,\mathscr{G}(t_2,x))$ holds, then the m-semiflow is said to be strict.

Definition 7. The m-semiflow $\mathscr{G}$ is said to be $\mathscr{B}(X)$-dissipative if there exists a set $B_0 \subset \mathscr{B}(X)$ such that for every $B \in \mathscr{B}(X)$ there exists $t_0 \in \mathbb{T}^+$ such that for all $t_0 \leq t \in \mathbb{T}^+$ we have $\mathscr{G}(t,B) \subset B_0$.

Definition 8. The m-semiflow $\mathscr{G}$ is said to be closed if for every $t \in \mathbb{T}^+$ from $u_n \to u$ and $w_n \to w$ (where both convergences must hold in strong topology of X) with $w_n \in \mathscr{G}(t,u_n)$ it follows that $w \in \mathscr{G}(t,u)$.

Definition 9. The m-semiflow $\mathscr{G}$ is said to be asymptotically compact if for every $B \in \mathscr{B}(X)$ and for all sequences $\{t_n\} \subset \mathbb{T}^+$ such that $t_n \to \infty$ and $\xi_n \in \mathscr{G}(t_n,B)$ for a subsequence we have $\xi_n \to \xi$ strongly in X with $\xi \in X$.

Definition 10. The m-semiflow $\mathscr{G}$ is said to be compact if for every $B \in \mathscr{B}(X)$ and for all $t > 0$ the set $\mathscr{G}(t,B)$ is relatively compact.

Definition 11. The set $\mathscr{A} \subset X$ is called a global attractor for an m-semiflow $\mathscr{G}$ if it is nonempty, compact, negatively semiinvariant (i.e., $\mathscr{A} \subset \mathscr{G}(t,\mathscr{A})$ for all $t \in \mathbb{T}^+$), and attracts all bounded sets in X, i.e. for all $B \in \mathscr{B}(X)$ we have

$$\lim_{t\to\infty} \text{dist}_X(\mathscr{G}(t,B),\mathscr{A}) = 0.$$

The global attractor is said to be invariant if for all $t \in \mathbb{T}^+$ we have $\mathscr{A} = \mathscr{G}(t,\mathscr{A})$.

We cite the following theorem on the existence of global attractor (see [5] Theorem 18, [25] Theorem 3 and Remark 8)

Theorem 3. *Let X be a Banach space. If an m-semiflow $\mathscr{G} : \mathbb{T}^+ \times X \to P(X)$ is $\mathscr{B}(X)$-dissipative, closed, and asymptotically compact, then is has a global attractor $\mathscr{A}$. If $\mathscr{G}$ is strict, then the attractor $\mathscr{A}$ is invariant.*

Note that the global attractor, if it exists, must be defined uniquely. Moreover it is the minimal closed set that attracts all sets from $\mathscr{B}(X)$ and it is the maximal bounded negatively semiinvariant set.

Lemma 1. *Let X be a Banach space. If an m-semiflow $\mathscr{G} : \mathbb{T}^+ \times X \to P(X)$ is $\mathscr{B}(X)$-dissipative and compact then it is asymptotically compact.*

Proof. Take $B \in \mathscr{B}(X)$ and $t_n \to \infty$. Choose $t > 0$. For n large enough we have $\mathscr{G}(t_n, B) \subset \mathscr{G}(t, \mathscr{G}(t_n - t, B))$. Again, for n large enough, from $\mathscr{B}(X)$-dissipativity we have $\mathscr{G}(t_n, B) \subset \mathscr{G}(t, B_0)$, and the assertion follows by compactness.

3 Problem Definition

Let V be a reflexive and separable Banach space, and let H be a Hilbert space. We consider an evolution triple $V \subset H \subset V^*$ with the embeddings being continuous, dense and compact. The embedding $V \subset H$ will be denoted by $i : V \to H$. The norm in V will be denoted by $\|\cdot\|$ while all other norms will be denoted by appropriate subscripts. Similarly, the duality pairing between V and V^* will be denoted by $\langle \cdot, \cdot \rangle$, while duality pairings between other spaces will be denoted by appropriate subscripts. Scalar product in H will be denoted by $(\cdot, \cdot)$. Let moreover U be another reflexive Banach space and let $\iota : V \to U$ be a linear, continuous and compact operator not identically equal to zero. The norm of ι will be denoted as $\|\iota\| := \|\iota\|_{\mathscr{L}(V;U)}$. We fix $p \geq 2$ and denote the adjoint exponent by q, i.e. $\frac{1}{p} + \frac{1}{q} = 1$. For a finite time interval I, the spaces of time dependent functions will be denoted, respectively, as $\mathscr{V}(I) = L^p(I;V)$, $\mathscr{V}^*(I) = L^q(I;V^*)$, $\mathscr{U}(I) = L^p(I;U)$. Moreover we use the notation $\mathscr{W}(I) = \{u \in \mathscr{V}(I) \,|\, u' \in \mathscr{V}^*(I)\}$, where u' is the derivative in the distributional sense. We will also use the notation $\mathscr{V}_{loc}(\mathbb{R}^+)$ (respectively $\mathscr{V}^*_{loc}(\mathbb{R}^+)$, $\mathscr{W}_{loc}(\mathbb{R}^+)$, $\mathscr{U}_{loc}(\mathbb{R}^+)$) for the spaces of functions that belong to $\mathscr{V}(0,T)$ (respectively $\mathscr{V}^*(0,T)$, $\mathscr{W}(0,T)$, $\mathscr{U}(0,T)$) for all $T > 0$.

We assume that $A : V \to V^*$ is a possibly nonlinear operator and $F : U \to 2^{U^*}$ is a multifunction. The detailed assumptions on the problem data are the following

$H(A)$

- (i) A is pseudomonotone.
- (ii) A is coercive in the sense that for all $v \in V$ we have $\langle Av, v \rangle \geq \alpha \|v\|^p$ with the constant $\alpha > 0$.
- (iii) A satisfies the growth condition $\|Av\|_{V^*} \leq a + b\|v\|^{p-1}$ for all $v \in V$ with $b > 0$ and $a \geq 0$.

$H(F)$

- (i) For every $u \in U$ the set $F(u)$ is nonempty, closed, and convex.
- (ii) F satisfies the growth condition $\|\xi\|_{U^*} \le c(1+\|u\|_U^{p-1})$ for all $u \in U$ and $\xi \in F(u)$ with the constant $c > 0$.
- (iii) Graph of F is a sequentially closed set in $(\text{strong} - U) \times (\text{weak} - U^*)$ topology.
- (iv) F satisfies the dissipativity condition $\langle \xi, u\rangle_{U^*\times U} \ge c_1 - c_2\|u\|_U^p$ for all $u \in U$ and $\xi \in F(u)$ with $c_1 \in \mathbb{R}$ and $0 \le c_2 < \frac{\alpha}{\|\iota\|^p}$.

(H_0) $f \in V^*$, $u_0 \in H$.

$H(U)$ For all finite time intervals I the Nemytskii mapping for ι denoted as $\bar{\iota} : M^{p,q}(I;V,V^*) \to \mathscr{U}(I)$ is compact.

Remark 1. Note that assumption $H(U)$ is motivated by Proposition 2 in [15]. It holds, for example, if we can find another Banach space Z such that $V \subset Z$ compactly, $Z \subset H$ continuously and there exists a linear and continuous mapping $\gamma : Z \to U$ such that for all $v \in V$ we have $\gamma v = \iota v$.

Remark 2. If the constant c_2 in $H(F)(iv)$ is negative, then the term with F has a dissipative nature. In such a case we can set $c_2 = 0$. Due to dissipativity of A it is possible that $c_2 > 0$, then the term with F is "excitatory" but the magnitude of c_2 is limited by the coercivity constant α.

The main problem under consideration is the following

$(\mathscr{P})$ Find $u \in \mathscr{W}_{loc}(\mathbb{R}^+)$ such that $u(0) = u_0$ and for a.e. $t > 0$ we have

$$u'(t) + Au(t) + \iota^* F(\iota u(t)) \ni f. \tag{1}$$

Note that we say that $u \in \mathscr{W}_{loc}(\mathbb{R}^+)$ solves (1) when there exists $\eta : \mathbb{R}^+ \to U^*$ such that for a.e. $t \in \mathbb{R}^+$ we have $\eta(t) \in F(\iota u(t))$ and $u'(t) + Au(t) + \iota^*\eta(t) = f$.

We fix $\tau > 0$ and we define the temporal semidiscretization of the problem $(\mathscr{P})$, where we use the implicit Euler scheme

$(\mathscr{P}_\tau)$ Find $\{u_\tau^k\}_{k=0}^\infty$ such that $u_\tau^0 = u_{0\tau}$, where $\{u_{0\tau}\} \subset H$ is a sequence such that $u_{0\tau} \to u_0$ strongly in H as $\tau \to 0$ and for all $k \in \mathbb{N}^+$ and $v \in V$ we have $u_\tau^k \in V$ and

$$\left(\frac{u_\tau^k - u_\tau^{k-1}}{\tau}, v\right) + \langle Au_\tau^k, v\rangle + \langle \eta_\tau^k, \iota u\rangle_{U^*\times U} = \langle f, v\rangle, \tag{2}$$

where $\eta_\tau^k \in F(\iota u_\tau^k)$.

We will use the notation $\mathbb{T}_\tau^+ = \{0, \tau, 2\tau, \ldots\}$. For a solution of Problem $(\mathscr{P}_\tau)$ we can construct piecewise constant and piecewise linear interpolants $\bar{u}_\tau, u_\tau : \mathbb{R}^+ \to H$

$$\bar{u}_\tau(t) = u_\tau^k, \quad u_\tau(t) = u_\tau^k \frac{t-(k-1)\tau}{\tau} + u_\tau^{k-1}\frac{k\tau - t}{\tau} \quad \text{for } t \in ((k-1)\tau, k\tau].$$

For simplicity we assume that $\bar{u}_\tau(0) = u_\tau^1$. Note that, denoting by $\bar{\eta}_\tau : \mathbb{R}^+ \to U^*$ the piecewise constant function equal to η_τ^k for $t \in ((k-1)\tau.k\tau]$, Eq. (2) is equivalent to the following equation in V^*

$$u_\tau'(t) + A\bar{u}_\tau(t) + \imath^*\bar{\eta}_\tau(t) = f, \tag{3}$$

valid for all $t > 0$.

4 Pseudomonotonicity of Nemytskii Operator for A

Before we pass to the analysis of Problem $(\mathscr{P})$ and its time discretized approximation we formulate and prove the auxiliary result on pseudomonotonicity of the Nemytskii operator for the operator A. This operator, denoted as $\mathfrak{A} : \mathscr{V}(0,T) \to \mathscr{V}^*(0,T)$ is defined by the expression $(\mathfrak{A}(u))(t) = A(u(t))$. The proof of the next theorem is inspired by the proof of Theorem 2 in [4] (compare Proposition 1 in [31] and Lemma 1 in [15]). Since our setup is different than in mentioned works, we present the proof here.

Theorem 4. *Fix $T > 0$. Let A satisfy $H(A)$ and let W be a Banach space such that the embedding $W \subset L^p(0,T;H)$ is compact. Then the Nemytskii operator $\mathfrak{A} : \mathscr{V}(0,T) \to \mathscr{V}^*(0,T)$ is W-pseudomonotone.*

Proof. Let $v_n \to v$ weakly in $\mathscr{V}(0,T)$ and $\|v_n\|_W$ be bounded with

$$\limsup_{n\to\infty} \langle \mathfrak{A}v_n, v_n - w\rangle_{\mathscr{V}(0,T)\times\mathscr{V}^*(0,T)} \le 0.$$

From the compactness of the embedding $W \subset L^p(0,T;H)$ it follows that $v_n \to v$ strongly in the latter space and, for a subsequence, $v_n(t) \to v(t)$ strongly in H for a.e. $t \in (0,T)$. The null set on which the convergence does not hold will be denoted by N. Define $\xi_n(t) = \langle Av_n(t), v_n(t) - v(t)\rangle$. We have, by $H(A)(ii)$ and $H(A)(iii)$

$$\xi_n(t) \ge \alpha\|v_n(t)\|^p - (a + b\|v_n(t)\|^{p-1})\|v(t)\| \ge \frac{\alpha}{2}\|v_n(t)\|^p - a\|v(t)\| - C\|v(t)\|^p, \tag{4}$$

with a constant $C > 0$. Let $D = \{t \in [0,T] \mid \liminf_{n\to\infty} \xi_n(t) < 0\}$. Obviously this is a measurable set. Assume that $m(D) > 0$. For $t \in D \setminus N$ the sequence $v_n(t)$ has, by (4) a subsequence which is bounded in V such that, for this subsequence, $\lim_{n\to\infty}\langle Av_n(t), v_n(t) - v\rangle < 0$. Again for a subsequence we have $v_n(t) \to v(t)$ weakly in V, where the limit is equal to $v(t)$ since we consider only $t \notin N$. By pseudomonotonicity of A we get $0 \le \liminf_{n\to\infty}\langle Av_n(t), v_n(t) - v(t)\rangle$, which is a contradiction. So, $m(D) = 0$, which means that $\liminf_{n\to\infty}\xi_n(t) \ge 0$ for a.e. $t \in (0,T)$. By (4) we can use the Fatou Lemma to get

$$0 \leq \int_0^T \liminf_{n\to\infty} \langle Av_n(t), v_n(t) - v(t)\rangle \, dt \leq \liminf_{n\to\infty} \langle \mathfrak{A}v_n, v_n - v\rangle_{\mathscr{V}^*(0,T)\times\mathscr{V}(0,T)}$$
$$\leq \limsup_{n\to\infty} \langle \mathfrak{A}v_n, v_n - v\rangle_{\mathscr{V}^*(0,T)\times\mathscr{V}(0,T)} \leq 0.$$

This means that $\lim_{n\to\infty} \int_0^T \xi_n(t)\,dt = 0$. Now note that $|\xi_n(t)| = \xi_n(t) + 2\xi_n^-(t)$, and we have $\lim_{n\to\infty} \xi_n^-(t) = 0$ for a.e. $t \in (0,T)$. By (4) we have $\xi_n^-(t) \leq a\|v(t)\| + C\|v(t)\|^p \in L^1(0,T)$. We can use the Fatou Lemma to get $\limsup_{n\to\infty} \int_0^T \xi_n^-(t)\,dt \leq 0$ and furthermore $\lim_{n\to\infty} \int_0^T \xi_n^-(t)\,dt = 0$. We deduce that $\xi_n \to 0$ in $L^1(0,T)$, and, for a subsequence, not renumbered, $\xi_n(t) \to 0$ a.e. $t \in (0,T)$. From (4) it follows that $v_n(t) \to v(t)$ weakly in V where the limit must be equal to $v(t)$ and the convergence must hold for the subsequence on which $v_n(t) \to v(t)$ strongly in H. Using the pseudomonotonicity of A it follows that

$$Av_n(t) \to Av(t) \text{ weakly in } V^* \text{ and } \langle Av_n(t), v_n(t)\rangle \to \langle Av(t), v(t)\rangle. \tag{5}$$

Choose $w \in \mathscr{V}(0,T)$. We have, using the Fatou Lemma once again

$$\liminf_{n\to\infty} \langle \mathfrak{A}v_n, v_n - w\rangle_{\mathscr{V}^*(0,T)\times\mathscr{V}(0,T)} \geq \int_0^T \liminf_{n\to\infty} \langle Av_n(t), v_n(t) - w(t)\rangle \, dt. \tag{6}$$

For subsequence of indices, which may be different for different t we have

$$\liminf_{n\to\infty} \langle Av_n(t), v_n(t) - w(t)\rangle = \lim_{k\to\infty} \langle Av_{n_k}(t), v_{n_k}(t) - w(t)\rangle.$$

From previous arguments, for another subsequence we have (5), and, by the uniqueness of the limit, the convergence must hold for whole n_k. Hence

$$\liminf_{n\to\infty} \langle Av_n(t), v_n(t) - w(t)\rangle = \langle Av(t), v(t) - w(t)\rangle,$$

which, together with (6), completes the proof.

5 Convergence of Semi-Discrete Scheme

In this section we formulate and prove the theorem on the convergence of semi-discrete scheme solutions to a solution of the time continuous problem. We first formulate the theorem, and then we present several auxiliary results useful in its proof.

Theorem 5. *Let $H(A), H(F), (H_0), H(U)$ hold. Then the Problem $(\mathscr{P})$ has a (possibly nonunique) solution. Moreover Problem $(\mathscr{P}_\tau)$ has a solution for every $\tau > 0$. This solution can be possibly nonunique in every time step. If $\tau \to 0$, then*

for each corresponding sequence of solutions to $(\mathscr{P}_\tau)$ *we can find a subsequence such that*

$$\bar{u}_\tau \to u \ \text{ weakly in } \ \mathscr{V}_{loc}(\mathbb{R}^+) \text{ and } \text{ weakly-* in } \ L^\infty_{loc}(\mathbb{R}^+;H), \tag{7}$$

$$\bar{\iota}\bar{u}_\tau \to \bar{\iota}u \ \text{ strongly in } \ \mathscr{U}_{loc}(\mathbb{R}^+), \tag{8}$$

$$u_\tau \to u \ \text{ weakly-* in } \ L^\infty_{loc}(\mathbb{R}^+;H), \tag{9}$$

$$u'_\tau \to u' \ \text{ weakly } \quad \mathscr{V}^*_{loc}(\mathbb{R}^+), \tag{10}$$

$$u_\tau(t) \to u(t) \ \text{ strongly in } \ H \ \text{ for all } \ t \in (0,\infty), \tag{11}$$

where u *solves the problem* (P).

It is enough to prove the above theorem for the finite time interval $(0,T)$ for a given and fixed T. Indeed, the solution u must belong to $C(0,T;H)$ and hence the value of this solution at the point T can be taken as the initial condition to construct another solution of length T. Then, by concatenation of these solutions it is possible to construct the solution on $(0,2T)$ and by repetition of this procedure on the whole positive semi-axis. Likewise, it is sufficient to prove the convergences (7)–(11) on the interval $(0,T)$ and use the diagonal argument to construct a subsequence of indexes such that (7)–(11) hold on the whole $\mathbb{R}^+$. In the sequel of this section for simplicity we assume that $\tau_n \to 0$ is a sequence such that $\frac{T}{\tau_n} := N_n$ is always a natural number. This assumption is made to avoid technical difficulties only. To simplify notation we will write τ, N in place of τ_n, N_n.

Lemma 2. *Let* $n \in \mathbb{N}$ *and* $k \in \{1,\dots,N\}$ *be given. Under assumptions* $H(A)$, $H(F)$, H_0 *and* $H(U)$, *for any* $\tau > 0$ *there exists* $u^k_\tau \in V$, *the solution to Problem* $(\mathscr{P}_\tau)$.

Proof. We rewrite equivalently (2) as follows:

$$\frac{1}{\tau}u^k_\tau + Au^k_\tau + \iota^* F(\iota u^k_\tau) \ni \frac{1}{\tau}u^{k-1}_\tau + f.$$

We show that, given $u^{k-1}_\tau \in V$, there exists u^k_τ that satisfies the above inclusion. To this end, we prove that the range of multifunction $V \ni v \to Lv = \frac{i^* i v}{\tau} + Av + \iota^* F(\iota v)$ is the whole space V^*. This will be done by a surjectivity theorem of Brézis (see, for instance, Theorem 1.3.70 in [13]). We need to show that L is coercive (in the sense that $\lim_{\|v\|\to\infty} \inf_{v^* \in Lv} \frac{\langle v^*, v\rangle}{\|v\|} = \infty$) and multivalued pseudomonotone. We will the use the fact that sum of multivalued pseudomonotone maps is multivalued pseudomonotone, cf. [13] Proposition 1.3.68. First, observe that the operator $\frac{i^* i}{\tau}$ satisfies conditions (i)–(iii) of Proposition 1 trivially. The fact that A satisfies these conditions follows from $H(A)(i)$ and $H(A)(iii)$. To show multivalued pseudomonotonicity of F observe that (i) of Proposition 1 follows from $H(F)(i)$, while (ii) is a consequence of $H(F)(ii)$. We prove (iii). Let $v_n \to v$ weakly in V and $v^*_n \to v^*$ weakly in V^* be such that $v^*_n \in \iota^* F(\iota v_n)$. Then $v^*_n = \iota^* \xi_n$ for certain $\xi_n \in F(\iota v_n)$. Compactness of ι implies that $\iota v_n \to \iota v$ strongly in U, and, by the growth condition $H(F)(ii)$, we have, for a subsequence, $\xi_n \to \xi$ weakly in U^*.

Now $H(F)(iii)$ implies that $\xi \in F(\imath v)$, and continuity of the adjoint mapping $\imath^*$ implies that $v_n^* = \imath^* \xi_n \to \imath^* \xi$ weakly in V^*, and $\imath^* \xi = v^*$. Moreover $v^* \in \imath^* F(\imath v)$ and $\langle v_n^*, v_n \rangle = \langle \xi_n, \imath v_n \rangle_{U^* \times U} \to \langle \xi, \imath v \rangle_{U^* \times U} = \langle v^*, v \rangle$, and the uniqueness of the limit implies that the convergence holds for the whole sequence. The assertion (iii) is proved. In order to show the coercivity of L we need to assume that $v^* \in Lv$ and estimate $\langle v^*, v \rangle$ from below. We have, with certain $\eta \in F(\imath v)$

$$\begin{aligned}\langle v^*, v \rangle &\geq \frac{\|v\|_H^2}{\tau} + \alpha \|v\|^p + \langle \eta, \imath v \rangle_{U^* \times U} \geq \frac{\|v\|_H^2}{\tau} + \alpha \|v\|^p - c_1 - c_2 \|\imath v\|^p \\ &\geq \frac{\|v\|_H^2}{\tau} + (\alpha - c_2 \|\imath\|^p) \|v\|^p - c_1,\end{aligned}$$

where we have used $H(A)(ii)$ and $H(A)(iv)$, whence the coercivity follows. The proof is complete.

The next result establishes estimates which are satisfied by the solutions of the semi-discrete problem.

Lemma 3. *Under assumptions $H(A), H(F), H(U)$, and H_0, if $\{u_\tau^k\}$ solve the Problem $(\mathscr{P}_\tau)$ then we have for all natural $m \geq 1$*

$$\|u_\tau^m\|_H^2 + \tau \sum_{k=1}^m \|u_\tau^k - u_\tau^{k-1}\|_H^2 + C_1 \tau \sum_{k=1}^m \|u_\tau^k\|^p \leq m\tau C_2 + \|u_\tau^0\|_H^2, \tag{12}$$

$$\|u_\tau^m\|_H^2 \leq C_3 + \|u_\tau^0\|_H^2 \frac{1}{(1 + C_4 \tau)^m}, \tag{13}$$

where the positive constants C_1, C_2, C_3, C_4 are independent of m, τ, u_τ^0.

Proof. The proof is standard. We take duality in (2) with u_τ^k and use the relation $\|a\|^2 - (a,b) = \frac{\|a\|^2 - \|b\|^2 + \|a-b\|^2}{2}$. Using $H(A)(ii)$ and $H(F)(iv)$ we obtain for $k \in \mathbb{N}^+$

$$\frac{1}{2\tau}(\|u_\tau^k\|_H^2 - \|u_\tau^{k-1}\|_H^2 + \|u_\tau^k - u_\tau^{k-1}\|_H^2) + (\alpha - c_2 \|\imath\|^p) \|u_\tau^k\|^p + c_1 \leq \langle f, u_\tau^k \rangle. \tag{14}$$

Now (12) follows by summing (14) from $k = 1$ to m and a straightforward computation.

To show (13) observe that from (14) it follows that, for some constants D_1, D_2 independent of k, τ, u_τ^0

$$\|u_\tau^k\|_H^2 + D_1 \tau \|u_\tau^k\|^p + \leq D_2 \tau + \|u_\tau^{k-1}\|_H^2.$$

Since $p \geq 2$, it follows that

$$\|u_\tau^k\|_H^2 (1 + D_3 \tau) \leq D_4 \tau + \|u_\tau^{k-1}\|_H^2,$$

where D_3, D_4 depend only on D_1, D_2, p and the norm of the embedding $i : V \to H$. By a simple induction it follows that (13) holds with $C_4 = D_3$ and $C_3 = \frac{D_4}{D_3}$.

Lemma 4. *Let $H(A),H(F),H(U),H_0$ hold and let $\tau>0$. The following estimates hold for all natural $m\geq 1$*

$$\tau\sum_{k=1}^{m}\|Au_\tau^k\|_{V^*}^q+\tau\sum_{k=1}^{m}\|\eta_\tau^k\|_{U^*}^q+\tau\sum_{k=1}^{m}\left\|\frac{u_\tau^k-u_\tau^{k-1}}{\tau}\right\|_{V^*}^q\leq M_1\|u_\tau^0\|_H^2+M_2m\tau, \tag{15}$$

where $\eta_\tau^k\in F(\iota u_\tau^k)$ are such that (2) holds and the constants $M_1,M_2>0$ depend only on the problem data (excluding τ,m,u_τ^0).

Proof. From the growth condition $H(A)(iii)$ we have

$$\sum_{k=1}^{m}\tau\|Au_\tau^k\|_{V^*}^q\leq\tau\sum_{k=1}^{m}(a+b\|u_\tau^k\|_{V^*}^{p-1})^q\leq 2^{q-1}m\tau a^q+2^{q-1}b^q\tau\sum_{k=1}^{m}\|u_\tau^k\|_{V^*}^p. \tag{16}$$

Next we observe that by the growth condition $H(F)(ii)$ we have

$$\tau\sum_{k=1}^{m}\|\eta_\tau^k\|_{U^*}^q\leq c^q\tau\sum_{k=1}^{m}(1+\|\iota u_\tau^k\|_U^{p-1})^q\leq c^q\tau m2^{q-1}+\|\iota\|^q c^q 2^{q-1}\tau\sum_{k=1}^{m}\|u_\tau^k\|^p. \tag{17}$$

In order to estimate the last term in (15), from (2) we obtain

$$\tau\sum_{k=1}^{m}\left\|\frac{u_\tau^k-u_\tau^{k-1}}{\tau}\right\|_{V^*}^q\leq\tau\sum_{k=1}^{m}\left\|f-Au_\tau^k-\iota^*\eta_\tau^k\right\|_{V^*}^q, \tag{18}$$

where $\eta_\tau^k\in F(\iota u_\tau^k)$. Moreover, we have

$$\tau\sum_{k=1}^{m}\left\|\frac{u_\tau^k-u_\tau^{k-1}}{\tau}\right\|_{V^*}^q\leq C\tau m\|f\|_{V^*}+C\tau\sum_{k=1}^{m}(\|Au_\tau^k\|_{V^*}^q+\|\iota\|^q\|\eta_\tau^k\|_{U^*}^q), \tag{19}$$

where $C>0$. Now the assertion (15) follows from the estimates (16), (17), (19), and (12) of Lemma 3. The proof is complete.

We formulate the following lemma.

Lemma 5. *Under assumptions $H(A),H(F),H(U),H_0$, the sequence $\{\bar u_\tau\}$ is bounded in $\mathscr{V}(0,T)\cap L^\infty(0,T;H)$ and the sequence $\{u_\tau\}$ is bounded in $C(0,T;H)$ with $\{u_\tau'\}$ bounded in $\mathscr{V}^*(0,T)$. Furthermore $\{\bar u_\tau\}$ is bounded in $BV^q(0,T;V^*)$. Finally $\{\mathfrak{A}\bar u_\tau\}$ and $\{\bar\eta_\tau\}$ are bounded in $\mathscr{V}^*(0,T)$ and $\mathscr{U}^*(0,T)$, respectively.*

Proof. It suffices to show the BV^q estimate since all the other estimates follow directly from Lemmata 3 and 4. The BV^q seminorm of $\bar u_\tau$ is given by

$$\|\bar u_\tau\|_{BV^q(0,T;V^*)}^q=\sum_{j=1}^{M_\tau}\|u_\tau^{m_\tau^j}-u_\tau^{m_\tau^{j-1}}\|_{V^*}^q,$$

and it is attained by the partition such that its vertices fall in the grid intervals indexed by $m_\tau^0, m_\tau^1, \ldots, m_\tau^{M_\tau-1}, m_\tau^{M_\tau}$, where $m_\tau^0 = 1$ and $m_\tau^{M_\tau} = N$. By the convexity of the function $h(s) = s^q$, we obtain

$$\begin{aligned}
\|\bar{u}_\tau\|^q_{BV^q(0,T;V^*)} &\leq \sum_{j=1}^{M_\tau} (m_\tau^{j-1} - m_\tau^j)^{q-1} \sum_{i=m_\tau^{j-1}+1}^{m_\tau^j} \|u_\tau^i - u_\tau^{i-1}\|^q_{V^*} \\
&\leq \sum_{j=1}^{M_\tau} (m_\tau^{j-1} - m_\tau^j)^{q-1} \tau^{q-1} \tau \sum_{i=m_\tau^{j-1}+1}^{m_\tau^j} \left\| \frac{u_\tau^i - u_\tau^{i-1}}{\tau} \right\|^q_{V^*} \\
&\leq (N\tau)^{q-1} \tau \sum_{i=2}^{N} \left\| \frac{u_\tau^i - u_\tau^{i-1}}{\tau} \right\|^q_{V^*},
\end{aligned}$$

and the assertion follows from (15).

The next Lemma establishes weak and weak-* limits of subsequences of constructed interpolants.

Lemma 6. *Under assumptions $H(A), H(F), H(U)$, and H_0, there exists $u \in \mathscr{W}(0,T)$ as well as $\zeta \in \mathscr{V}^*(0,T)$, $\eta \in \mathscr{U}^*(0,T)$ and a subsequence of indices such that for this subsequence (still denoted by the index τ), we have*

$$\bar{u}_\tau \to u \ \text{ weakly in } \ \mathscr{V} \ \text{ and weakly-* in } \ L^\infty(0,T;H), \tag{20}$$

$$u_\tau \to u \ \text{ weakly-* in } \ L^\infty(0,T;H), \tag{21}$$

$$u'_\tau \to u' \ \text{ weakly in } \ \mathscr{V}^*, \tag{22}$$

$$\bar{\imath}\bar{u}_\tau \to \bar{\imath}u \ \text{ strongly in } \ \mathscr{U}, \tag{23}$$

$$\mathfrak{A}\bar{u}_n \to \zeta \ \text{ weakly in } \ \mathscr{V}^*, \tag{24}$$

$$\bar{\eta}_\tau \to \eta \ \text{ weakly in } \ \mathscr{U}^*. \tag{25}$$

Proof. The fact that the limits of appropriate subsequences exist follows directly from Lemmata 3–5, and $H(U)$. It only suffices to prove that limits of u_τ and $\bar{u}_\tau$ coincide. This is done in a standard way by showing the estimate on $\|u_\tau - \bar{u}_\tau\|_{\mathscr{V}^*}$. By the direct calculation we have

$$\begin{aligned}
\|u_\tau - \bar{u}_\tau\|^q_{\mathscr{V}^*} &= \sum_{k=1}^{N} \int_{(k-1)\tau}^{k\tau} \left\| u_\tau^k - u_\tau^{k-1} - \frac{u_\tau^k - u_\tau^{k-1}}{\tau} (t - (k-1)\tau) \right\|^q_{V^*} dt \\
&\leq \frac{\tau^q}{q+1} \sum_{k=1}^{N} \tau \left\| \frac{u_\tau^k - u_\tau^{k-1}}{\tau} \right\|^q_{V^*}.
\end{aligned}$$

By the estimate (15), it follows that $u_\tau - \bar{u}_\tau \to 0$ in $\mathscr{V}^*(0,T)$ as $\tau \to 0$ and therefore the limits of two sequences must coincide.

Theorem 6. *Under assumptions $H(A), H(F), H(U)$, and H_0, the function u obtained in Lemma 6 solves Problem $(\mathscr{P})$.*

Proof. First we show that u satisfies the initial condition. From (21) and (22), it follows, by Corollary 4 of [35], that

$$u_\tau \to u \ \text{ strongly in } \ C(0,T;V^*), \tag{26}$$

and furthermore $u_\tau(0) \to u(0)$ strongly in V^*. Since $u_\tau(0) = u_{0\tau}$ and $u_{0\tau} \to u_0$ strongly in H, from the uniqueness of the limit, it follows that $u(0) = u_0$.

Let us observe that (3) can be equivalently rewritten as the following equality that holds in $\mathscr{V}^*(0,T)$ for all τ

$$u'_\tau + \mathfrak{A}\bar{u}_\tau + \bar{\iota}^*\bar{\eta}_\tau = f. \tag{27}$$

We can pass to the limit in (27) and find

$$u' + \zeta + \bar{\iota}^*\eta = f. \tag{28}$$

To conclude the proof we must verify that $\zeta = \mathfrak{A}u$ and $\eta(t) \in F(\iota u(t))$ for a.e. t.

First, we verify this last assertion. Since $\bar{\iota}\bar{u}_\tau \to \bar{\iota}u$ strongly in $\mathscr{U}(0,T)$, then, moreover, for a subsequence, $\iota\bar{u}_\tau(t) \to \iota u(t)$ strongly in U for a.e. $t \in (0,T)$ with $\|\iota u_\tau(t)\|_U \le a(t)$ for certain function $a \in L^p(0,T)$. Moreover $\bar{\eta}_\tau \to \eta$ weakly in $L^1(0,T;U^*)$, with $\bar{\eta}_\tau(t) \in F(\iota\bar{u}_\tau(t))$ for a.e. $t \in (0,T)$.

Growth condition $H(F)(ii)$ implies that for a.e. $t \in (0,T)$ we have

$$\|\bar{\eta}_\tau(t)\|_{U^*} \le c(1 + \|\iota\bar{u}_\tau(t)\|_U^{p-1}) \le c(1 + a(t)^{p-1}).$$

We can use Theorem 2 to deduce that

$$\eta(t) \in \overline{\text{conv}}\ \text{weak} - \limsup_{\tau\to 0}\{\bar{\eta}_\tau(t)\} \subset \overline{\text{conv}}\ \text{weak} - \limsup_{\tau\to 0}\ F(\iota\bar{u}_\tau(t))$$

for a.e. $t \in (0,T)$. The strong convergence $\iota\bar{u}_\tau(t) \to \iota u(t)$ and the assumptions $H(F)(i)$ and $H(F)(iii)$ imply that

$$\eta(t) \in \overline{\text{conv}}F(\iota u(t)) = F(\iota u(t))$$

for a.e. $t \in (0,T)$ and the assertion is proved.

Subsequently, we verify that $\zeta = \mathfrak{A}u$. We use the relation

$$\limsup_{\tau\to 0}\langle \mathfrak{A}\bar{u}_\tau, \bar{u}_\tau - u\rangle_{\mathscr{V}^*(0,T)\times\mathscr{V}(0,T)} = \limsup_{\tau\to 0}\langle f - u'_n - \bar{\iota}^*\bar{\eta}_n, \bar{u}_n - u\rangle_{\mathscr{V}^*(0,T)\times\mathscr{V}(0,T)},$$

whence it follows that

$$\limsup_{\tau\to 0}\langle \mathfrak{A}\bar{u}_\tau, \bar{u}_\tau - u\rangle_{\mathscr{V}^*(0,T)\times\mathscr{V}(0,T)} = \limsup_{\tau\to 0}\langle u'_\tau, u - \bar{u}_\tau\rangle_{\mathscr{V}^*(0,T)\times\mathscr{V}(0,T)}.$$

By the direct calculation, we have

$$\langle u'_\tau, u-\bar{u}_\tau\rangle_{\mathscr{V}^*(0,T)\times\mathscr{V}(0,T)} = \langle u'_\tau, u\rangle_{\mathscr{V}^*(0,T)\times\mathscr{V}(0,T)} - \langle u'_\tau, \bar{u}_\tau\rangle_{\mathscr{V}^*(0,T)\times\mathscr{V}(0,T)}$$
$$= \langle u'_\tau, u\rangle_{\mathscr{V}^*(0,T)\times\mathscr{V}(0,T)} + \frac{1}{2}(\|u_{0\tau}\|_H^2 - \|u_\tau(T)\|_H^2) - \frac{1}{2}\sum_{k=1}^{N}\|u_\tau^k - u_\tau^{k-1}\|_H^2.$$

Hence

$$\limsup_{\tau\to 0}\langle u'_\tau, u-\bar{u}_\tau\rangle_{\mathscr{V}^*(0,T)\times\mathscr{V}(0,T)}$$
$$\le \langle u', u\rangle_{\mathscr{V}^*(0,T)\times\mathscr{V}(0,T)} + \frac{1}{2}(\|u_0\|_H^2 - \liminf_{\tau\to 0}\|u_\tau(T)\|_H^2).$$

Finally, we observe that $\|u_\tau(T)\|_H$ is bounded and therefore we may assume that for a subsequence $u_\tau(T)\to w$ weakly in H with $w\in H$. It follows from (26) that $u_\tau(T)\to u(T)$ strongly in V^*. We conclude that $w=u(T)$ and the convergence holds for the whole subsequence for which the assertion of Lemma 6 holds. From the weak lower-semicontinuity of the norm, we have

$$\limsup_{\tau\to 0}\langle u'_\tau, u-\bar{u}_\tau\rangle_{\mathscr{V}^*(0,T)\times\mathscr{V}(0,T)} \le \frac{\|u_0\|_H^2 - \|u(T)\|_H^2}{2} + \langle u', u\rangle_{\mathscr{V}^*(0,T)\times\mathscr{V}(0,T)} = 0,$$

which gives

$$\limsup_{\tau\to 0}\langle \mathfrak{A}\bar{u}_\tau, \bar{u}_\tau - u\rangle_{\mathscr{V}^*(0,T)\times\mathscr{V}(0,T)} \le 0. \tag{29}$$

By Theorems 1 and 4 it follows that the Nemytskii operator $\mathfrak{A}$ is $M^{p,q}(0,T;V,V^*)$-pseudomonotone. We conclude that

$$\langle \mathfrak{A}u, u-y\rangle_{\mathscr{V}^*(0,T)\times\mathscr{V}(0,T)} \le \liminf_{\tau\to 0}\langle \mathfrak{A}\bar{u}_\tau, \bar{u}_\tau - y\rangle_{\mathscr{V}^*(0,T)\times\mathscr{V}(0,T)},$$

for all $y\in\mathscr{V}$. The assertion follows taking, respectively, $y=u+w$ and $y=u-w$, where $w\in\mathscr{V}^*$.

We formulate the following theorem.

Theorem 7. *Let $\varepsilon>0$. Under assumptions $H(A), H(F), H(U)$, and H_0 for the convergent subsequence established by Lemma 6, we have $u_\tau\to u$ weakly in $L^p(\varepsilon,T;V)$ and $u_\tau(t)\to u(t)$ strongly in H for all $t\in[0,T]$.*

Proof. Let us first estimate

$$\int_{(k-1)\tau}^{k\tau}\|\bar{u}_\tau(t)-u_\tau(t)\|^p\,dt = \frac{\tau}{p+1}\|u_\tau^k - u_\tau^{k-1}\|^p \le \frac{2^{p-1}\tau}{p+1}(\|u_n^k\|^p + \|u_n^{k-1}\|^p).$$

Fix $\varepsilon > 0$. Let $K(\tau,\varepsilon)$ be the smallest index such that $K(\tau,\varepsilon)\tau > \varepsilon$. We have

$$\|\bar{u}_\tau - u_\tau\|^p_{L^p(\varepsilon,T;V)} \leq \sum_{i=K(\tau,\varepsilon)}^{N} \frac{2^{p-1}\tau}{p+1}(\|u^k_\tau\|^p + \|u^{k-1}_\tau\|^p)$$
$$\leq \frac{2^p\tau}{p+1} \sum_{i=K(\tau,\varepsilon)-1}^{N} \|u^k_\tau\|^p.$$

Since $\tau \to 0$, for small enough τ we have $K(\tau,\varepsilon) - 1 \geq 1$ and, from Lemma 5 we obtain that u_τ is bounded in $L^p(\varepsilon,T;V)$. It follows that $u_\tau \to u$ weakly in this space and the convergence must hold for the whole subsequence for which Lemma 5 holds. Moreover, by Lemma 5, u'_n is bounded in $\mathscr{V}^*(0,T)$ and in $\mathscr{V}^*(\varepsilon,T)$. Hence, by the Aubin–Lions compactness theorem it follows that $u_\tau \to u$ strongly in $L^p(\varepsilon,T;H)$. From arbitrariness of ε it follows that $u_\tau(t) \to u(t)$ strongly in H for a.e. $t \in (0,T)$, where we do not need to pass to subsequence since we already know that $u_\tau \to u$ strongly in $C(0,T;V^*)$. To prove the strong convergence for all $t \in [0,T]$ observe that $u_\tau(0) = u_{0\tau} \to u_0 = u(0)$ strongly in H. Pick $t > 0$. There exists $\varepsilon \in (0,t)$ such that $u_\tau(\varepsilon) \to u(\varepsilon)$ strongly in H. Subtracting Eq. (3) from (1), taking the duality with $\bar{u}_\tau - u$ and integrating over the interval (ε,t), we obtain

$$\langle u'_\tau - u', \bar{u}_\tau - u\rangle_{\mathscr{V}^*(\varepsilon,t)\times\mathscr{V}(\varepsilon,t)} + \langle \mathfrak{A}\bar{u}_\tau - \mathfrak{A}u, \bar{u}_\tau - u\rangle_{\mathscr{V}^*(\varepsilon,t)\times\mathscr{V}(\varepsilon,t)} \tag{30}$$
$$+ \langle \bar{\eta}_\tau - \eta, \bar{\imath}\bar{u}_\tau - \bar{\imath}u\rangle_{\mathscr{U}^*(\varepsilon,t)\times\mathscr{U}(\varepsilon,t)} = 0.$$

We have $\bar{\imath}\bar{u}_\tau \to \bar{\imath}u$ strongly in $\mathscr{U}(0,T)$ and moreover in $\mathscr{U}(\varepsilon,t)$. Furthermore, by Lemma 5 the sequence $\bar{\eta}_n$ is bounded in $\mathscr{U}^*(0,T)$ and moreover in $\mathscr{U}^*(\varepsilon,t)$. Hence, we get

$$\lim_{\tau\to 0}\langle \bar{\eta}_\tau - \eta, \bar{\imath}\bar{u}_\tau - \bar{\imath}u\rangle_{\mathscr{U}^*(\varepsilon,t)\times\mathscr{U}(\varepsilon,t)} = 0. \tag{31}$$

Using (31) in (30), we have

$$\lim_{\tau\to 0}\langle u'_\tau - u' + \mathfrak{A}\bar{u}_\tau, \bar{u}_\tau - u\rangle_{\mathscr{V}^*(\varepsilon,t)\times\mathscr{V}(\varepsilon,t)} = 0. \tag{32}$$

Now we need to prove that

$$\limsup_{\tau\to 0}\langle \mathfrak{A}\bar{u}_\tau, \bar{u}_\tau - u\rangle_{\mathscr{V}^*(\varepsilon,t)\times\mathscr{V}(\varepsilon,t)} \leq 0.$$

This follows analogously as the proof of (29) in Theorem 6. The only delicate step in the proof is the computation of $-\langle u'_\tau, \bar{u}_n\rangle_{\mathscr{V}^*(\varepsilon,t)\times\mathscr{V}(\varepsilon,t)}$. We have

$$-\langle u'_\tau, \bar{u}_n\rangle_{\mathscr{V}^*(\varepsilon,t)\times\mathscr{V}(\varepsilon,t)} = \frac{\|u_\tau(\varepsilon)\|_H^2 - \|u_\tau(t)\|_H^2}{2} + \int_\varepsilon^t (u'_\tau(s), u_\tau(s) - \bar{u}_\tau(s))\,ds$$
$$\leq \frac{\|u_\tau(\varepsilon)\|_H^2 - \|u_\tau(t)\|_H^2}{2}, \tag{33}$$

where the inequality holds since, as the simple calculation shows, the integrand is nonpositive for all $s \in (0,T)$. Now, from Theorem 4, it follows that

$$0 \leq \liminf_{\tau\to 0} \langle \mathfrak{A}\bar{u}_\tau, \bar{u}_\tau - u\rangle_{\mathscr{V}^*(\varepsilon,t)\times\mathscr{V}(\varepsilon,t)},$$

so

$$\langle \mathfrak{A}\bar{u}_\tau, \bar{u}_\tau - u\rangle_{\mathscr{V}^*(\varepsilon,t)\times\mathscr{V}(\varepsilon,t)} \to 0.$$

Thus (32) implies

$$\lim_{\tau\to 0} \langle u'_\tau - u', \bar{u}_\tau - u_\tau + u_\tau - u\rangle_{\mathscr{V}^*(\varepsilon,t)\times\mathscr{V}(\varepsilon,t)} = 0. \tag{34}$$

The last equation can be reformulated as

$$\lim_{\tau\to 0} \left(\frac{1}{2}(\|u_\tau(t) - u(t)\|_H^2 - \|u_\tau(\varepsilon) - u(\varepsilon)\|_H^2) + \right.$$
$$\left. + \int_\varepsilon^t (u'_\tau(s), \bar{u}_\tau(s) - u_\tau(s))\,ds - \langle u', \bar{u}_\tau - u_\tau\rangle_{\mathscr{V}^*(\varepsilon,t)\times\mathscr{V}(\varepsilon,t)} \right) = 0.$$

Since $u_\tau(\varepsilon) \to u(\varepsilon)$ strongly in H and $\bar{u}_\tau - u_\tau \to 0$ weakly in $\mathscr{V}(0,T)$ and also weakly in $\mathscr{V}(\varepsilon,t)$ we can write

$$0 = \lim_{\tau\to 0} \left(\frac{1}{2}\|u_\tau(t) - u(t)\|_H^2 + \int_\varepsilon^t (u'_\tau(s), \bar{u}_\tau(s) - u_\tau(s))\,ds \right).$$

Since the integrand in above expression must be positive, analogously to (33) we get

$$0 \geq \frac{1}{2} \limsup_{\tau\to 0} \|u_\tau(t) - u(t)\|_H^2,$$

and it follows that $\|u_\tau(t) - u(t)\|_H \to 0$, which concludes the proof.

Now, Theorem 5 follows from Lemma 6 and Theorems 6 and 7.

6 Global Attractor for Time Continuous Problem

First we prove an a priori estimate on the solutions of Problem $(\mathscr{P})$.

Lemma 7. *Let u solve Problem $(\mathscr{P})$ with the initial condition $u_0 \in H$. Let moreover η be a corresponding selection of $F(\imath u(t))$. We have the following estimates*

$$\|u(t)\|_H^2 + C_1 \int_0^t \|u(s)\|^p \, dt \leq C_2 t + \|u_0\|_H^2, \tag{35}$$

$$\|u(t)\|_H^2 \leq \|u_0\|_H^2 e^{-C_3 t} + C_4, \tag{36}$$

$$\int_0^t \|u'(s)\|_{V^*}^q \, ds + \int_0^t \|Au(s)\|_{V^*}^q \, ds + \int_0^t \|\eta(s)\|_{U^*}^q \, ds \leq C_5 t + C_6 \|u_0\|_H^2. \tag{37}$$

valid for all $t \geq 0$ with the constants $C_1, \ldots, C_6 > 0$.

Proof. We take the duality in (1) with $u(t)$ and use $H(A)(ii)$ and $H(F)(iv)$. We get for a.e. $t \geq 0$ with arbitrary $\varepsilon > 0$ and the positive constant $C(\varepsilon)$

$$\frac{1}{2}\frac{d}{dt}\|u(t)\|_H^2 + (\alpha - c_2\|\imath\|^p - \varepsilon)\|u(t)\|^p \leq C(\varepsilon)\|f\|_{V^*}^q - c_1.$$

Now we can choose $\varepsilon = \frac{1}{2}(\alpha - c_2\|\imath\|^p)$ and (35) follows by direct integration. A version of Young inequality $a^2 \leq a^p + D_1$ valid for $a \geq 0$ with a constant $D_1 > 0$ and the inequality $\|u\|_H \leq \|i\|\|v\|$, where $\|i\| := \|i\|_{\mathscr{L}(V;H)}$ is the norm of the embedding operator yields

$$\frac{d}{dt}\|u(t)\|_H^2 + D_2\|u(t)\|_H^2 \leq D_3,$$

for a.e. $t > 0$ with constants $D_2, D_3 > 0$. The application of the Gronwall inequality gives (36). Now (37) follows in a straightforward way by the application of growth conditions $H(A)(iii)$, $H(F)(ii)$, the estimate (35), and Eq. (1) for the estimate on the derivative.

We can define the multivalued map $\mathscr{G} : \mathbb{R}^+ \times H \to P(H)$ by

$$\mathscr{G}(t, u_0) = \{u(t) \mid u \text{ solves Problem } (\mathscr{P}) \text{ with initial condition } u_0\}.$$

Obviously, $\mathscr{G}$ is a strict m-semiflow. We continue to investigate the properties of $\mathscr{G}$

Lemma 8. *The m-semiflow $\mathscr{G}$ is $\mathscr{B}(H)$-dissipative.*

Proof. The assertion follows directly from the estimate (36).

Lemma 9. *The m-semiflow $\mathscr{G}$ is closed.*

Proof. Choose $T \geq 0$. Let $u_{n0} \to u_0$ strongly in H and $\mathscr{G}(T, u_{n0}) \ni w_n \to w$ strongly in H. We must show that $w \in \mathscr{G}(T, u_0)$. There exist the trajectories u_n such that

$u_n(0) = u_{n0}$ and $u_n(T) = w_n$. Denote the corresponding selections of $F(\imath u_n(t))$ by η_n. From Lemma 7 it follows that $\{u_n\}$ is bounded in $\mathscr{W}(0,T)$ and from the following estimate valid for any interval $(a,b) \subset (0,T)$

$$\|u_n(b) - u_n(a)\|_{V^*}^q \le \left\| \int_a^b u_n'(t)\,dt \right\|_{V^*}^q \le \int_a^b \|v'(t)\|_{V^*}^q\,dt (b-a)^{\frac{q}{p}}$$
$$\le \int_a^b \|v'(t)\|_{V^*}^q\,dt T^{\frac{q}{p}},$$

we deduce that $\{u_n\}$ is bounded in $M^{p,q}(0,T;V,V^*)$. Hence, by $H(U)$, for a subsequence, we have

$$u_n \to u \text{ weakly in } \mathscr{V}(0,T) \text{ and strongly in } L^p(0,T;H), \tag{38}$$
$$\bar{\imath} u_n \to \bar{\imath} u \text{ strongly in } \mathscr{U}(0,T), \tag{39}$$
$$u_n' \to u' \text{ weakly in } \mathscr{V}^*(0,T). \tag{40}$$

Moreover, by Lemma 7 it follows that for certain $\xi \in \mathscr{V}^*(0,T)$ and $\eta \in \mathscr{U}^*(0,T)$ we have

$$\mathfrak{A} u_n \to \xi \text{ weakly in } \mathscr{V}^*(0,T), \tag{41}$$
$$\eta_n \to \eta \text{ weakly in } \mathscr{U}^*(0,T). \tag{42}$$

Since for $v \in V$ and $t \in [0,T]$ we have

$$(u_n(t) - u(t), v) = (u_{n0} - u(0), v) + \int_0^t \langle u_n'(s) - u'(s), v\rangle\,ds, \tag{43}$$

we deduce the equality

$$\int_0^T (u_n(t) - u(t), v)\,dt = T(u_{n0} - u(0), v) + \int_0^T \int_0^t \langle u_n'(s) - u'(s), v\rangle\,ds\,dt$$

valid for $v \in V$, whence by (38) and (40) it follows that $u_{n0} \to u(0)$ weakly in V^* and hence $u(0) = u_0$. From (43) written for $t = T$ and (40) it follows that $u_n(T) \to u(T)$ weakly in V^* and we deduce that $w = u(T)$. It remains to show that u solves Problem $(\mathscr{P})$. We can pass to the limit in (1) and obtain

$$u'(t) + \xi(t) + \imath^* \eta(t) = f(t) \text{ a.e. } t \in (0,T).$$

By (39) we have, for a subsequence, $\imath u_n(t) \to \imath u(t)$ and $\|\imath u_n(t)\|_U \le a(t)$ with $a \in L^p(0,T)$ a.e. $t \in (0,T)$ and by (42) we have $\eta_n \to \eta$ weakly in $L^1(0,T;U^*)$. Growth condition $H(F)(ii)$ implies that

$$\|\eta_n(t)\|_{U^*} \le c(1 + \|\imath u_n(t)\|_U^{p-1}) \le c(1 + a(t)^{p-1})$$

for a.e. $t \in (0,T)$. We are in position to use Theorem 2, whence

$$\eta(t) \in \overline{\text{conv}} \text{ weak} - \limsup_{n\to\infty} \{\eta_n(t)\} \subset \overline{\text{conv}} \text{ weak} - \limsup_{n\to\infty} F(\iota u_n(t))$$

for a.e. $t \in (0,T)$. The assumption $H(F)(iii)$ implies that

$$\eta(t) \in \overline{\text{conv}}\, F(\iota u(t))$$

for a.e. $t \in (0,T)$ and by $H(F)(i)$ we deduce that $\eta(t) \in F(\iota u(t))$ a.e. $t \in (0,T)$. It remains to prove that $\xi(t) = Au(t)$ a.e. $t \in (0,T)$, or, in other words $\xi = \mathfrak{A}u$. Since u_n is bounded in $\mathscr{W}(0,T)$ which embeds in $L^p(0,T;H)$ compactly, the Nemytskii operator $\mathfrak{A}$ is $\mathscr{W}(0,T)$-pseudomonotone by Theorem 4. Moreover $u_n \to u$ weakly in $\mathscr{V}(0,T)$. Next, we calculate

$$\begin{aligned}
&\int_0^T \langle Au_n(t), u_n(t) - u(t)\rangle\, dt = \int_0^T \langle f(t) - u_n'(t) - \iota^* \eta_n(t), u_n(t) - u(t)\rangle\, dt \\
&= \int_0^T \langle f(t), u_n(t) - u(t)\rangle\, dt - \int_0^T \langle \eta_n(t), \iota u_n(t) - \iota u(t)\rangle_{U^*\times U}\, dt \\
&+ \int_0^T \langle u_n'(t), u(t)\rangle\, dt - \frac{\|u_n(T)\|_H^2 - \|u_n(0)\|_H^2}{2}.
\end{aligned}$$

We can pass to the limit in all terms in the last equality which gives us

$$\lim_{n\to\infty} \int_0^T \langle Au_n(t), u_n(t) - u(t)\, dt = \int_0^T \langle u'(t), u(t)\rangle\, dt - \frac{\|u(T)\|_H^2 - \|u(0)\|_H^2}{2} = 0.$$

Hence, taking $w = v - z$ for $z \in \mathscr{V}(0,T)$ in the Definition 2 we get

$$\int_0^T \langle Au(t), z(t)\rangle\, dt \le \liminf_{n\to\infty} \int_0^T \langle Au_n(t), z(t)\rangle\, dt.$$

Taking $-z$ in place of z it follows that $\mathfrak{A}u_n \to \mathfrak{A}u$ weakly in $\mathscr{V}^*$, and the proof is complete.

Lemma 10. *The m-semiflow $\mathscr{G}$ is compact.*

Proof. Let $t > 0$ and let $\{u_{0n}\}$ be a sequence, bounded in H. Moreover, let $w_n \in \mathscr{G}(t, u_{0n})$. We must show that $\{w_n\}$ is relatively compact, that is, it contains a subsequence that converges strongly in H. By (35) the sequence w_n is bounded in H and hence, for a subsequence, $w_n \to w$ weakly in H for some $w \in H$. We will continue the argument for this subsequence. There exist the trajectories u_n such that $u_n(0) = u_{n0}$ and $u_n(t) = w_n$. The corresponding selections of $F(\iota u_n(t))$ are denoted by η_n. By Lemma 7 the sequence $\{u_n\}$ is bounded in $\mathscr{W}(0,2t)$ and $M^{p,q}(0,2t;V,V^*)$ and hence, analogously as in the proof of Lemma 9, for a subsequence, we have

$$u_n \to u \text{ weakly in } \mathscr{V}(0,2t), \tag{44}$$

$$\eta_n \to \eta \text{ weakly in } \mathscr{U}^*(0,2t), \tag{45}$$

$$u_n \to u \text{ strongly in } L^p(0,2t;H), \tag{46}$$

$$\bar{\iota} u_n \to \bar{\iota} u \text{ strongly in } \mathscr{U}(0,2t), \tag{47}$$

$$u_n(s) \to u(s) \text{ strongly in } H \text{ a.e. } s \in (0,2t). \tag{48}$$

If $u_n(t) \to u(t)$ strongly in H, then the proof is complete. Suppose this is not the case. We take duality in (1) written for u_n with $u_n(t)$ and integrate over the interval $(0,s)$ for $s \in (0,2t)$. We obtain

$$\begin{aligned}&\frac{1}{2}\|u_{0n}\|_H^2 - \int_0^s \langle Au_n(r), u_n(r)\rangle\, dr\\ &= \frac{1}{2}\|u_n(s)\|_H^2 + \int_0^s \langle \eta_n(r), \iota u_n(r)\rangle_{U^*\times U}\, dr - \int_0^r \langle f, u_n(r)\rangle\, dr.\end{aligned}$$

We introduce the auxiliary functions $V_n, V : (0,2t) \to \mathbb{R}$

$$V_n(s) = \frac{1}{2}\|u_{0n}\|_H^2 - \int_0^s \langle Au_n(r), u_n(r)\rangle\, dr.$$

$$V(s) = \frac{1}{2}\|u(s)\|_H^2 + \int_0^s \langle \eta(r), \iota u(r)\rangle_{U^*\times U}\, dr - \int_0^r \langle f, u(r)\rangle\, dr.$$

By $H(A)(ii)$ the functions V_n are nonincreasing in time. Moreover, $V_n(s) \to V(s)$ for a.e. $s \in (0,2t)$. Let $t_l \searrow t$ and $s_r \nearrow t$ be the sequences such that this convergence holds. We have for all n, r, l

$$V_n(s_r) \geq V_n(t) \geq V_n(t_l)$$

passing with $n \to \infty$ we get

$$V(s_r) \geq \limsup_{n\to\infty} V_n(t) \geq \liminf_{n\to\infty} V_n(t) \geq V(t_l).$$

Now we pass with l, r to infinity. Since $\mathscr{W}(0,2t) \subset C(0,2t;H)$ the mapping $s \to \|u(s)\|_H$ is continuous, moreover V is continuous and we obtain

$$V(t) \geq \limsup_{t\to\infty} V_n(t) \geq \liminf_{n\to\infty} V_n(t) \geq V(t),$$

and hence $V_n(t) \to V(t)$ as $n \to \infty$ and it follows that $\|w_n\| = \|u_n(t)\|_H \to \|u(t)\|_H$. Since we already know that $w_n \to w$ weakly in H, from the fact that H is a Hilbert space it follows that $w_n \to w$ strongly in H and the proof of compactness is complete.

In view of Theorem 3 and Lemma 1, using Lemmata 8–10, we have shown the following

Theorem 8. *Under assumptions $H(A), H(F), (H_0), H(U)$, the m-semiflow $\mathscr{G}$ associated with the solution of Problem $(\mathscr{P})$ has a global attractor $\mathscr{A}$ which is moreover invariant.*

7 Global Attractor for the Time Discrete Problem

First, as a simple consequence of Lemma 3, we formulate the following corollary:

Lemma 11. *There exists $R > 0$ such that for any $B \in \mathscr{B}(H)$ and any $\tau > 0$ we can find $n_0(B,\tau) \in \mathbb{N}$ such that if $u_\tau^0 = u_0 \in B$ and $\{u_\tau^k\}_{k=1}^\infty$ solve $(\mathscr{P}_\tau)$ with the initial condition u_τ^0 then for every $n \geq n_0$ we have $\|u_\tau^n\|_H \leq R$.*

We can define the multifunction $\mathscr{G}_\tau : \mathbb{T}_\tau^+ \times H \to P(H)$ in the following way:

- $\mathscr{G}_\tau(0,u) = \{u\}$,
- $\mathscr{G}_\tau(\tau,u)$ it the set of all u_τ^k that solve Problem $(\mathscr{P}_\tau)$ with $u_\tau^{k-1} = u$,
- $\mathscr{G}_\tau(k\tau,u) = \mathscr{G}(\tau,\mathscr{G}((k-1)\tau,u))$ for $k \in \{2,3,\dots\}$.

Obviously, $\mathscr{G}_\tau$ defines a strict m-semiflow, and from Lemma 11 it follows that $\mathscr{G}_\tau$ is $\mathscr{B}(H)$-dissipative.

Lemma 12. *The m-semiflow $\mathscr{G}_\tau$ is closed.*

Proof. Fix $\tau > 0$ and $n \in \mathbb{N}^+$. Assume that $u_{\tau m}^0 \to u_\tau^0$ strongly in H as $m \to \infty$ and $\mathscr{G}_\tau(n\tau, u_{\tau m}^0) \ni u_{\tau m}^n \to u_\tau^n$ strongly in H. We must prove that $u_\tau^n \in \mathscr{G}_\tau(n\tau, u_\tau^0)$. There exist the corresponding discrete trajectories $u_{\tau m}^k \in \mathscr{G}_\tau(\tau, u_{\tau m}^{k-1})$ for $k \in \{1,\dots,n\}$ with η_τ^k in (2) denoted as $\eta_{\tau m}^k \in F(\iota u_{\tau m}^k)$. From (12) it follows that for all $k \in \{1,\dots,n\}$ we have the bound $\|u_{\tau m}^k\| \leq C$, where C is a constant dependent on τ, n but independent of k, m. We are therefore in position to construct the subsequence of indices, not renumbered, such that for all $k \in \{1,\dots,n\}$ we have $u_{\tau m}^k \to u_\tau^k$ weakly in V as $m \to \infty$. Hence, $\iota u_{\tau m}^k \to \iota u_\tau^k$ strongly in U and $u_{\tau m}^k \to u_\tau^k$ strongly in H. From growth conditions $H(A)(iii)$ and $H(F)(ii)$ for the subsequences, which are not renumbered, we have

$$Au_{\tau m}^k \to \xi_\tau^k \text{ weakly in } V^* \qquad \text{and} \qquad \bar{\eta}_{\tau m}^k \to \eta_\tau^k \text{ weakly in } U^*.$$

We can pass to the limit in (2) written for m and we get

$$\left(\frac{u_\tau^k - u_\tau^{k-1}}{\tau}, v\right) + \langle \xi_\tau^k, v\rangle + \langle \eta_\tau^k, \iota u\rangle_{U^*\times U} = \langle f, v\rangle.$$

By the sequential strong–weak closedness of the graph of F (cf. $H(F)(iii)$) it follows that $\eta_\tau^k \in F(\iota u_\tau^k)$. It remains to prove that $\xi_\tau^k = Au_\tau^k$. We will use the pseudomonotonicity of A. Let us calculate

$$\begin{aligned}
&\langle Au_{\tau m}^k, u_{\tau m}^k - u_\tau^k\rangle \\
&= \langle f, u_{\tau m}^k - u_\tau^k\rangle - \langle \eta_{\tau m}^k, \iota u_{\tau m}^k - \iota u_\tau^k\rangle_{U^*\times U} - \left(\frac{u_{\tau m}^k - u_{\tau m}^{k-1}}{\tau}, u_{\tau m}^k - u_\tau^k\right).
\end{aligned}$$

Clearly, $\lim_{m\to\infty}\langle Au^k_{\tau m}, u^k_{\tau m}-u^k_\tau\rangle = 0$. Hence for any $z\in V$ we have

$$\langle Au^k_\tau, z\rangle \le \liminf_{m\to\infty}\langle Au^k_{\tau m}, u^k_{\tau m}-u^k_\tau+z\rangle = \liminf_{m\to\infty}\langle Au^k_{\tau m}, z\rangle = \langle \xi^k_\tau, z\rangle,$$

and the proof is complete.

Lemma 13. *The m-semiflow $\mathscr{G}_\tau$ is compact.*

Proof. Let $B\in\mathscr{B}(H)$ and a sequence $u^k_{m\tau}\in\mathscr{G}(k\tau,B)$ for some natural $k>0$. By (12) we have $\|u^k_{m\tau}\|\le C$, where the constant $C>0$ depends on τ, which is fixed. Hence, for a subsequence, $u^k_{\tau m}\to w$ as $m\to\infty$ weakly in V, for certain $w\in V$. Moreover $u^k_{\tau m}\to w$ strongly in H, and the proof is complete.

In view of Theorem 3 and Lemma 1, since by Lemma 11 the m-semiflow $\mathscr{G}$ is $\mathscr{B}(H)$-dissipative, by Lemma 12 it is closed, and by Lemma 13 it is compact, we have shown the following Theorem

Theorem 9. *Let $\tau>0$ be given. Under assumptions $H(A),H(F),(H_0),H(U)$, the m-semiflow $\mathscr{G}_\tau$ associated with the solution of Problem $(\mathscr{P}_\tau)$ has a global attractor $\mathscr{A}_\tau$ which is moreover invariant.*

8 Upper-Semicontinuous Convergence of Attractors

We formulate and prove the following theorem on upper-semicontinuous convergence of the semi-discrete attractors to the time continuous attractor. For simplicity we set $\tau_n=\frac{1}{n}$, but the result remains valid for any sequence $\tau_n\to 0$.

Theorem 10. *Under assumptions $H(A),H(F),(H_0),H(U)$ we have*

$$\lim_{n\to\infty}\operatorname{dist}_H(\mathscr{A}_{\tau_n},\mathscr{A})=0.$$

Proof. Let $\mathscr{K}=\overline{\bigcup_{n=1}^\infty \mathscr{A}_{\tau_n}}^H$.

Step 1. Compactness of $\mathscr{K}$. We need to show that $\mathscr{K}$ is a compact set in H. To this end choose a subsequence of τ_n, which we will denote by the same index and $y_n\in\mathscr{A}_{\tau_n}$. It is enough to show that $\{y_n\}$ has a subsequence that converges strongly in H. From the attractor invariance we have $y_n\in\mathscr{G}_{\tau_n}(1,z_n)$ for some $z_n\in\mathscr{A}_{\tau_n}$. Since, by Lemma 11 the ball $B(0,R)$ is absorbing for all τ, it follows that $\|z_n\|_H\le R$ and, for a subsequence, not renumbered, we have $z_n\to z$ weakly in H. We can construct discrete trajectories $\{z^j_{\tau_n}\}_{j=0}^{2n}$ such that $z^0_{\tau_n}=z_n$ and $z^n_{\tau_n}=y_n$. Corresponding piecewise constant and piecewise linear interpolants will be denoted by $\bar z_{\tau_n}$ and z_{τ_n}. Moreover we denote the piecewise constant selection of $F(\iota\bar z_{\tau_n}(t))$, such that (3) holds, by $\bar\eta_{\tau_n}$ with $\bar\eta_{\tau_n}(t)=\eta^j_{\tau_n}$ for $t\in((j-1)\tau_n,j\tau_n]$.

Observe that we have $\|\bar{z}_{\tau_n}\|^p_{\mathscr{V}(0,2)} = \tau\sum_{j=1}^{2n}\|z^j_{\tau_n}\|^p$, and hence, by Lemma 3 the sequence $\bar{z}_{\tau_n}$ is bounded in $\mathscr{V}(0,2)$. Now let us compute for $j \geq 2$ the integral

$$\int_{(j-1)\tau_n}^{j\tau_n} \|z_{\tau_n}(t) - \bar{z}_{\tau_n}(t)\|^p \, dt = \|z^j_{\tau_n} - z^{j-1}_{\tau_n}\|^p \frac{\tau_n}{p+1}.$$

Hence, provided $n \geq 2$, we have

$$\begin{aligned}\|z_{\tau_n} - \bar{z}_{\tau_n}\|^p_{\mathscr{V}(\frac{1}{2},2)} &\leq \frac{\tau_n}{p+1}\sum_{j=[\frac{n}{2}]+1}^{2n}\|z^j_{\tau_n} - z^{j-1}_{\tau_n}\|^p \leq \frac{\tau_n 2^p}{p+1}\sum_{j=[\frac{n}{2}]}^{2n}\|z^j_{\tau_n}\|^p \\ &\leq \frac{2^p}{p+1}\|\bar{z}_{\tau_n}\|^p_{\mathscr{V}(0,2)}.\end{aligned}$$

This means that z_{τ_n} is bounded in $\mathscr{V}(\frac{1}{2},2)$. Now we estimate the derivative

$$\begin{aligned}\|z'_{\tau_n}\|^q_{\mathscr{V}^*(0,2)} &= \tau_n\sum_{j=1}^{2n}\left\|\frac{z^j_{\tau_n} - z^{j-1}_{\tau_n}}{\tau_n}\right\|^q_{V^*} = \tau_n\sum_{j=1}^{2n}\|f - Az^j_{\tau_n} - \iota^*\eta^j_{\tau_n}\|^q_{V^*} \\ &\leq 3^{q-1}\tau_n\sum_{j=1}^{2n}\left(\|f\|^q_{V^*} + \|Az^j_{\tau_n}\|^q_{V^*} + \|\iota\|^q\|\eta^j_{\tau_n}\|^q_{U^*}\right) \\ &\leq 3^{q-1}\Bigg(2\|f\|^q_{V^*} + 2^q a^q + 2^{q-1}b^q\tau_n\sum_{j=1}^{2n}\|z^j_{\tau_n}\|^p \\ &\quad + \|\iota\|^q 2^q c^q + \|\iota\|^{p+q}2^{q-1}c^q\tau_n\sum_{j=1}^{2n}\|z^j_{\tau_n}\|^p_V\Bigg),\end{aligned}$$

where we used the growth conditions $H(A)(iii)$ and $H(F)(ii)$. Hence, from boundedness of $\bar{z}_{\tau_n}$ in $\mathscr{V}(0,2)$ it follows that z'_{τ_n} is bounded in $\mathscr{V}^*(0,2)$. Finally, we need to establish the estimate on $\|\bar{z}_{\tau_n}\|_{BV^q(0,2;V^*)}$. Let $m^0_n = 1, \ldots, m^{M_n}_n = 2n$ be the indices of intervals in which the endpoints of the partition realizing the q-variation seminorm are contained. We have

$$\begin{aligned}\|\bar{z}_{\tau_n}\|^q_{BV^q(0,2;V^*)} &= \sum_{j=1}^{M_n}\|z^{m^j_n}_{\tau_n} - z^{m^{j-1}_n}_{\tau_n}\|^q_{V^*} \\ &\leq \sum_{j=1}^{M_n}\left((m^j_n - m^{j-1}_n)^{q-1}\sum_{i=m^{j-1}_n+1}^{m^j_n}\|z^i_{\tau_n} - z^{i-1}_{\tau_n}\|^q_{V^*}\right) \\ &\leq 2^{q-1}n^{q-1}\tau_n^{q-1}\tau_n\sum_{i=2}^{2n}\left\|\frac{z^i_{\tau_n} - z^{i-1}_{\tau_n}}{\tau_n}\right\|^q_{V^*} \leq 2^{q-1}\|z'_{\tau_n}\|^q_{\mathscr{V}^*(0,2)},\end{aligned}$$

and the required estimate follows. Now since z'_{τ_n} is bounded in $\mathscr{V}^*(0,2)$ and z_{τ_n} is bounded in $\mathscr{V}(\frac{1}{2},2)$, by the Aubin–Lions compactness theorem it follows that there exists $z \in L^p\left(\frac{1}{2},2;H\right)$, such that for a subsequence, we have $z_{\tau_n} \to z$ strongly in $L^p\left(\frac{1}{2},2;H\right)$ and moreover $z_{\tau_n}(t) \to z(t)$ strongly in H for a.e. $t \in \left(\frac{1}{2},2\right)$. Should this convergence hold for all $t \in \left(\frac{1}{2},2\right)$, then we would have the desired convergence $z_{\tau_n}(1) = y_n \to z(1)$. Note that, since $\mathscr{W}\left(\frac{1}{2},2\right) \subset C\left(\frac{1}{2},2;H\right)$ it must be that $z \in C\left(\frac{1}{2},2;H\right)$ and moreover $z_{\tau_n}(t) \to z(t)$ weakly in H for all $t \in \left[\frac{1}{2},2\right]$. Using $H(U)$ it follows that since $\bar{z}_{\tau_n}$ is bounded in $M^{p,q}(0,2;V,V^*)$, then there exists $\bar{z} \in \mathscr{V}(0,2)$ such that $\imath\bar{z}_{\tau_n} \to \imath\bar{z}$ strongly in $\mathscr{U}(0,2)$ and $\bar{z}_{\tau_n} \to \bar{z}$ weakly in $\mathscr{V}(0,2)$. Moreover, from the growth condition $H(F)(ii)$ and the bound on $\bar{z}_{\tau_n}$ in $\mathscr{V}(0,2)$ it follows that $\bar{\eta}_{\tau_n}$ is bounded in $L^q(0,2;U^*)$ and hence, for a subsequence we have $\bar{\eta}_{\tau_n} \to \bar{\eta}$ weakly in $L^q(0,2;U^*)$. Taking the duality in (3) with $\bar{z}_{\tau_n}(t)$ and integrating from 0 to t we get

$$\begin{aligned}&\frac{1}{2}\|z_n\|_H^2 - \int_0^t \langle A\bar{z}_{\tau_n}(s), \bar{z}_{\tau_n}(s)\rangle \, ds - \int_0^t \langle z'_{\tau_n}(s), \bar{z}_{\tau_n}(s) - z_{\tau_n}(t)\rangle \, ds \\ &= \frac{1}{2}\|z_{\tau_n}(t)\|_H^2 + \int_0^t \langle \bar{\eta}_{\tau_n}(s), \imath\bar{z}_{\tau_n}(s)\rangle_{U^*\times U} \, ds - \int_0^t \langle f, \bar{z}_{\tau_n}(s)\rangle \, ds.\end{aligned}$$

We introduce the discrete energy functions $V_n : [0,2] \to \mathbb{R}$ as

$$V_n(t) = \frac{1}{2}\|z_n\|_H^2 - \int_0^t \langle A\bar{z}_{\tau_n}(s), \bar{z}_{\tau_n}(s)\rangle \, ds - \int_0^t \langle z'_{\tau_n}(s), \bar{z}_{\tau_n}(s) - z_{\tau_n}(s)\rangle \, ds,$$

and the continuous energy function $V : (\frac{1}{2},2) \to \mathbb{R}$ as

$$V(t) = \frac{1}{2}\|z(t)\|_H^2 + \int_0^t \langle \bar{\eta}(s), \imath\bar{z}(s)\rangle_{U^*\times U} \, ds - \int_0^t \langle f, \bar{z}(s)\rangle \, ds.$$

By a straightforward computation we have $\langle z'_{\tau_n}(s), \bar{z}_{\tau_n}(s) - z_{\tau_n}(s)\rangle \geq 0$ for all $s \in [0,2]$. Due to this fact and by $H(A)(ii)$ we have that V_n are nonincreasing in time. Since

$$\int_0^t \langle \bar{\eta}_{\tau_n}(s), \imath\bar{z}_{\tau_n}(s)\rangle_{U^*\times U} - \langle f, \bar{z}_{\tau_n}(s)\rangle \, ds \to \int_0^t \langle \bar{\eta}(s), \imath\bar{z}(s)\rangle_{U^*\times U} - \langle f, \bar{z}(s)\rangle \, ds,$$

for all $t \in [0,2]$ and $z_{\tau_n}(t) \to z(t)$ for a.e. $t \in (\frac{1}{2},2)$ we have $V_n(t) \to V(t)$ for a.e. $t \in (\frac{1}{2},2)$. Let $t_l \searrow 1$ and $s_r \nearrow 1$ be the sequences such that this convergence holds. We have for all n,r,l

$$V_n(s_r) \geq V_n(1) \geq V_n(t_l)$$

passing with $n \to \infty$ we get

$$V(s_r) \geq \limsup_{n\to\infty} V_n(1) \geq \liminf_{n\to\infty} V_n(1) \geq V(t_l).$$

Now we pass with l, r to infinity and we obtain, by continuity of V

$$V(1) \geq \limsup_{n\to\infty} V_n(1) \geq \liminf_{n\to\infty} V_n(1) \geq V(1),$$

and hence $V_n(1) \to V(1)$ as $n \to \infty$ and it follows that $\|y_n\|_H \to \|z(1)\|_H$. Since we already know that $y_n = z^n_{\tau_n} = z_{\tau_n}(1) \to z(1)$ weakly in H, it follows that $y_n \to z(1)$ strongly in H and the proof of compactness is complete.

Step 2. Convergence of attractors. From the triangle inequality we have, for all $t \geq 0$

$$\mathrm{dist}_H(\mathscr{A}_{\tau_n}, \mathscr{A}) \leq \mathrm{dist}_H(\mathscr{A}_{\tau_n}, \mathscr{G}(t, \mathscr{K})) + \mathrm{dist}_H(\mathscr{G}(t, \mathscr{K}), \mathscr{A}).$$

Now fix $\varepsilon > 0$. Since $\mathscr{K}$ is bounded and $\mathscr{A}$ is the attractor for $\mathscr{G}$ we are able to find sufficiently large t_0 such that for all $t \geq t_0$ we have $\mathrm{dist}_H(\mathscr{G}(t, \mathscr{K}), \mathscr{A}) \leq \frac{\varepsilon}{2}$. We will show that for all $t \in \mathbb{N}$ we have

$$\lim_{n\to\infty} \mathrm{dist}_H(\mathscr{A}_{\tau_n}, \mathscr{G}(t, \mathscr{K})) = 0.$$

Suppose that this is not true. Then for some $t_0 \in \mathbb{N}^+$ and $\varepsilon_0 > 0$ and for subsequence of indices we can construct $y_n \in \mathscr{A}_{\tau_n}$ such that $\mathrm{dist}_H(y_n, \mathscr{G}(t_0, \mathscr{K})) \geq \varepsilon_0$. By invariance of semi-discrete attractors there exist $z_n \in \mathscr{A}_{\tau_n} \subset \mathscr{K}$ such that $y_n \in \mathscr{G}_{\tau_n}(t_0, z_n)$. Since $\mathscr{K}$ is compact, for a subsequence we have $z_n \to \hat{z}$ strongly in H with $\hat{z} \in \mathscr{K}$. We can construct discrete trajectories $\{z^j_{\tau_n}\}_{j=0}^{t_0 n}$ such that $z^0_{\tau_n} = z_n$ and $z^{t_0 n}_{\tau_n} = y_n$. Piecewise constant and piecewise linear interpolants corresponding to these discrete trajectories are denoted, respectively, by $\bar{z}_{\tau_n}$ and z_{τ_n}. From Theorem 5 there exists the solution z to Problem $(\mathscr{P})$ with the initial condition $\hat{z}$ such that $y_n = z_{\tau_n}(t_0) \to z(t_0)$ strongly in H. But $z(t_0) \in \mathscr{G}(t_0, \mathscr{K})$, and we have the contradiction with the fact that $\mathrm{dist}_H(y_n, \mathscr{G}(t_0, \mathscr{K})) \geq \varepsilon_0$. The proof is complete.

9 Examples

In this section we discuss the examples of problem data that satisfy the formulated assumptions and we give some applications of the problems analyzed in this section.

Assumption $H(U)$ Let $\Omega \subset \mathbb{R}^d$ be a bounded and open domain with Lipschitz boundary. Let moreover $\Gamma_D \subset \partial\Omega$ be a relatively open set of positive boundary measure and let $M \in \mathbb{N}$. Then we can set $V = \{v \in W^{1,p}(\Omega; \mathbb{R}^M) \mid$

$v = 0$ on $\Gamma_D\}$ and $H = L^2(\Omega;\mathbb{R}^M)$. We have two possible choices of U. One choice is $U = H$ and $\iota = i$. Obviously ι is compact, and by Theorem 1 so is the Nemytskii operator $\bar{\iota} : M^{p,q}(0,T;V,V^*) \to L^p(0,T;H)$. For the second choice let $\Gamma_C \subset \partial\Omega$ be a relatively open set of positive boundary measure disjoint with Γ_D. Let moreover $U = L^p(\Gamma_C;\mathbb{R}^M)$. Now $\iota : V \to U$ will be the trace operator. This operator is compact. Indeed, define $Z = W^{\delta,p}(\Omega;\mathbb{R}^M) \cap V$, equipped with $W^{\delta,p}$ topology, where $\frac{1}{p} < \delta < 1$. The embedding $V \subset Z$ is compact, and the trace $\gamma : Z \to U$ is linear and continuous, and hence ι is compact. Moreover, by Theorem 1 the Nemytskii operator $\bar{\iota} : M^{p,q}(0,T;V,V^*) \to L^p(0,T;U)$ is also compact.

Assumption $H(A)$ Detailed discussion and examples of pseudomonotone operators can be found in Chap. 27 of [38]. Roughly speaking, such operators are the sums of two terms: the first one given in divergence form, containing the highest order space derivatives that satisfy the so-called Leray–Lions conditions, and the second one being strongly continuous, depending only on lower order space derivatives. Detailed discussion of these assumptions is presented in Chap. 2 of [33].

Assumption $H(F)$ A notable example of multifunction that satisfies $H(F)(i)$ and $H(F)(iii)$ is the Clarke subdifferential of a locally Lipschitz functional. If $J : U \to \mathbb{R}$ is locally Lipschitz, then its Clarke directional derivative at the point $x \in U$ and in the direction $v \in U$ is defined as

$$J^0(x;v) = \limsup_{z \to x, \lambda \to 0^+} \frac{J(z+\lambda u) - J(z)}{\lambda},$$

and the Clarke subdifferential $\partial J : U \to P(U^*)$ is given by

$$\partial J(x) = \{\xi \in U^* \mid J^0(x,v) \geq \langle \xi, v \rangle_{U^* \times U} \text{ for all } v \in X\}.$$

For the properties of the Clarke subdifferential see [10].

To study another example involving Clarke subdifferential let $j : \mathbb{R}^M \to \mathbb{R}$, and let $U = L^p(\Gamma;\mathbb{R}^M)$ where either $\Gamma = \Omega$ or $\Gamma \subset \partial\Omega$ is a relatively open set, and $\Omega \subset \mathbb{R}^d$ is open and bounded with Lipschitz boundary. We assume that

(i) j is locally Lipschitz,
(ii) for all $s \in \mathbb{R}^M$ and $\xi \in \partial j(s)$ we have $|\xi| \leq c(1+|s|)$, with the constant $c > 0$.

Under these assumptions the mapping $N : U \to 2^{U^*}$ defined as $N(u) = \{\xi \in L^q(\Gamma;\mathbb{R}^M) \,|\, \xi(x) \in \partial j(u(x))$ a.e. $x \in \Gamma\}$ satisfies $H(F)(i)$–(iii) (see [3] for the proof). Partial differential inclusions with multivalued terms having the form of the Clarke subdifferential are known as *subdifferential inclusions* or *hemivariational inequalities*. They are used, for example, to model contact conditions in solid mechanics [29]. Existence of solutions for the first order in time parabolic problems with this type of multivalued terms was studied, for example, in [22, 27, 28].

Applications First order in time partial differential inclusions with multivalued terms that have strong–weak closed graph, but are not necessarily Hausdorff

continuous, can be used to model nonconvex semipermeability problems (Chap. 6.1 in [27]), temperature control problems (see [37]), problems modeling combustion in porous media or conduction of electrical impulses in nerve axon (Chap. 4.1.3 in [2]) and in climatology to model the energy balance of the Earth surface (see [2] Chap. 4.1.4). The analysis presented here can be applied to all mentioned problems, which shows that they and their time semidiscretizations possess global attractors, and that the semidiscrete attractors approximate, in the upper-semicontinuous sense, the global attractors of original problems.

Acknowledgements The research was supported by a Marie Curie International Research Staff Exchange Scheme Fellowship within the seventh European Community Framework Programme under Grant Agreement no. 2011-295118, the project Polonium "Mathematical and Numerical Analysis for Contact Problems with Friction" 2014/2015 between the Jagiellonian University and Université de Perpignan Via Domitia, the National Science Center of Poland under grant no. N N201 604640, the International Project co-financed by the Ministry of Science and Higher Education of Republic of Poland under grant no. W111/7.PR/2012, and the National Science Center of Poland under Maestro Advanced Project no. DEC-2012/06/A/ST1/00262.

References

1. Babin, A.V., Vishik, M.I.: Maximal attractors of semigroups corresponding to evolution differential equations. Mat. Sb. (N.S.) **126**, 397–419 (1985)
2. Balibrea, F., Caraballo, T., Kloeden, P.E., Valero, J.: Recent developments in dynamical systems: three perspectives. Int. J. Bifurc. Chaos **20**, 2591–2636 (2010)
3. Barboteu, M., Bartosz, K., Kalita, P., Ramadan, A.: Analysis of a contact problem with normal compliance, finite penetration and nonmonotone slip dependent friction. Commun. Contemp. Math. **16**, 1350016 [29 pages] (2014). doi: 10.1142/S0219199713500168
4. Berkovits, J., Mustonen, V.: Monotonicity methods for nonlinear evolution equations. Nonlinear Anal. Theory Methods Appl. **27**, 1397–1405 (1996)
5. Caraballo, T., Martin-Rubio, P., Robinson, J.C.: A comparison between two theories for multivalued semiflows and their asymptotic behavior. Set-Valued Anal. **11**, 297–322 (2003)
6. Carl, S.: Existence and comparison results for noncoercive and nonmonotone multivalued elliptic problems. Nonlinear Anal. Theory Methods Appl. **65**, 1532–1546 (2006)
7. Chepyzhov, V.V., Vishik, M.I.: Trajectory attractors for reaction-diffusion systems. Topol. Methods Nonlinear Anal. **7**, 49–76 (1996)
8. Cholewa, J., Dłotko, T.: Global Attractors in Abstract Parabolic Problems. Cambridge University Press, Cambridge (2000)
9. Chueshov, I., Lasiecka, I.: Von Karman Evolution Equations. Springer, New York (2010)
10. Clarke, F.H.: Optimization and Nonsmooth Analysis. SIAM, Philadelphia (1990)
11. Coti Zelati, M., Tone, F.: Multivalued attractors and their approximation: applications to the Navier-Stokes equations. Numer. Math. **122**, 421–441 (2012)
12. Denkowski, Z., Migórski, S., Papageorgiou, N.S.: An Introduction to Nonlinear Analysis: Theory. Kluwer Academic Publishers, Boston (2003)
13. Denkowski, Z., Migórski, S., Papageorgiou, N.S.: An Introduction to Nonlinear Analysis: Applications. Kluwer Academic Publishers, Boston (2003)
14. Gorban, N.V., Kapustyan, O.V., Kasyanov, P.O.: Uniform trajectory attractor for non-autonomous reaction-diffusion equations with Carathéodory's nonlinearity. Nonlinear Anal. Theory Methods Appl. **98**, 13–26 (2014)

15. Kalita, P.: Convergence of Rothe scheme for hemivariational inequalities of parabolic type. Int. J. Numer. Anal. Model. **10**, 445–465 (2013)
16. Kalita, P., Łukaszewicz, G.: Global attractors for multivalued semiflows with weak continuity properties. Nonlinear Anal. Theory Methods Appl.**101**, 124–143 (2014)
17. Kasyanov, P.O.: Multivalued dynamics of solutions of an autonomous differential-operator inclusion with pseudomonotone nonlinearity. Cybern. Syst. Anal. **47**, 800–811 (2011)
18. Kasyanov, P.O.: Multivalued dynamics of solutions of autonomous operator differential equations with pseudomonotone nonlinearity. Math. Notes **92**, 205–218 (2012)
19. Kasyanov, P.O., Toscano, L., Zadoyanchuk, N.V.: Long-time behaviour of solutions for autonomous evolution hemivariational inequality with multi-dimensional "reaction-displacement" law. Abstr. Appl. Anal. **2012**, Article ID 450984 (2012)
20. Kasyanov, P.O., Toscano, L., Zadoyanchuk, N.V.: Regularity of weak solutions and their attractors for a parabolic feedback control problem. Set-Valued Var. Anal. **21**, 271–282 (2013)
21. Kloeden, P.E., Valero, J.: Attractors for set-valued partial differential equations under discretization. IMA J. Numer. Anal. **29**, 690–711 (2009)
22. Liu, Z.: A class of parabolic hemivariational inequalities. Appl. Math. Mech. **21**, 1045–1052 (2000)
23. Łukaszewicz, G.: On the existence of the exponential attractor for a planar shear flow with Tresca's friction condition. Nonlinear Anal.-Real **14**, 1585–1600 (2013)
24. Malek, J., Nečas, J.: A finite-dimensional attractor for three-dimensional flow of incompressible fluids. J. Differ. Equ. **127**, 498–518 (1996)
25. Melnik, V.S., Valero, J.: On attractors of multivalued semiflows and differential inclusions. Set-Valued Anal. **6**, 83–111 (1998)
26. Melnik, V.S., Valero, J.: Addendum to "On attractors of multivalued semiflows and differential inclusions" [Set-Valued Anal. 6, 83–111 (1998)]. Set-Valued Anal. **16**, 507–509 (2008)
27. Miettinen, M., Panagiotopoulos, P.D.: On parabolic hemivariational inequalities and applications. Nonlinear Anal. Theory Methods Appl. **35**, 885–915 (1999)
28. Migórski, S., Ochal, A.: Boundary hemivariational inequality of parabolic type. Nonlinear Anal. Theory Methods Appl. **57**, 579–596 (2004)
29. Migórski, S., Ochal, A., Sofonea, M.: Nonlinear Inclusions and Hemivariational Inequalities. Models and Analysis of Contact Problems. Advances in Mechanics and Mathematics, vol. 26. Springer, New York (2013)
30. Miranville, A., Zelik, S.: Attractors for dissipative partial differential equations in bounded and unbounded domains. In: Evolutionary Equations: Handbook of Differential Equations, vol. IV, pp. 103–200. Elsevier, North-Holland/Amsterdam (2008)
31. Papageorgiou, N.S.: On the existence of solutions for nonlinear parabolic problems with nonmonotone discontinuities. J. Math. Anal. Appl. **205**, 434–453 (1997)
32. Robinson, J.C.: Infinite-Dimensional Dynamical Systems. Cambridge University Press, Cambridge (2001)
33. Roubíček, T.: Nonlinear Partial Differential Equations with Applications. Birkhäuser, Basel/Boston/Berlin (1990)
34. Sell, G.R.: Global attractors for the three dimensional Navier-Stokes equations. J. Dyn. Differ. Equ. **8**, 1–33 (1996)
35. Simon, J.: Compact sets in the space $L^p(0,T;B)$. Ann. Mat. Pura Appl. **146**, 65–96 (1987)
36. Temam, R.: Infinite-Dimensional Dynamical Systems in Mechanics and Physics, 2nd edn. Springer, New York (1997)
37. Wang, G., Yang, X.: Finite difference approximation of a parabolic hemivariational inequalities arising from temperature control problem. Int. J. Numer. Anal. Mod. **7**, 108–124 (2010)
38. Zeidler, E.: Nonlinear Functional Analysis and Its Applications. Nonlinear Monotone Operators, vol. II/B. Springer, New York (1990)

39. Zgurovsky, M.Z., Kasyanov, P.O.: Multivalued dynamics of solutions for autonomous operator differential equations in strongest topologies. In: Continuous and Distributed Systems, pp. 149–162. Springer, Heidelberg (2014)
40. Zhong, C.K., Yang, M.H., Sun, C.Y.: The existence of global attractors for the norm-to-weak continuous semigroup and application to the nonlinear reaction-diffusion equations. J. Differ. Equ. **223**, 367–399 (2006)

Passive Control of Singularities by Topological Optimization: The Second-Order Mixed Shape Derivatives of Energy Functionals for Variational Inequalities

Günter Leugering, Jan Sokołowski, and Antoni Żochowski

Abstract A class of nonsmooth shape optimization problems for variational inequalities is considered. The variational inequalities model elliptic boundary value problems with the Signorini type unilateral boundary conditions. The shape functionals are given by the first order shape derivatives of the elastic energy. In such a way the singularities of weak solutions to elliptic boundary value problems can be characterized. An example in solid mechanics is given by the Griffith's functional, which is defined in plane elasticity to measure SIF, the so-called stress intensity factor, at the crack tips. Thus, topological optimization can be used for passive control of singularities of weak solutions to variational inequalities.

The Hadamard directional differentiability of metric the projection onto the positive cone in fractional Sobolev spaces is employed to the topological sensitivity analysis of weak solutions of nonlinear elliptic boundary value problems. The first order shape derivatives of energy functionals in the direction of specific velocity fields depend on the solutions to variational inequalities in a subdomain. A domain decomposition technique is used in order to separate the unilateral boundary conditions and the energy asymptotic analysis.

The topological derivatives of nonsmooth integral shape functionals for variational inequalities are derived. Singular geometrical domain perturbations in an elastic body Ω are approximated by regular perturbations of bilinear forms in

G. Leugering
Chair of Applied Mathematics 2, Friedrich-Alexander University Erlangen-Nürnberg, Cauerstrasse 11, 91058 Erlangen, Germany
e-mail: leugering@math.fau.de

J. Sokołowski (✉)
Laboratoire de Mathématiques, Institut Elie Cartan de *Nancy*. campus Aiguillettes - BP 70239. 54506 VANDOEUVRE-LES-*NANCY* CEDEX, France
e-mail: Jan.Sokolowski@univ-lorraine.fr

A. Żochowski
Systems Research Institute of the Polish Academy of Sciences, ul. Newelska 6, 01-447 Warszawa, Poland
e-mail: Antoni.Zochowski@ibspan.waw.pl

J.-B. Hiriart-Urruty et al. (eds.), *Advances in Mathematical Modeling, Optimization and Optimal Control*, Springer Optimization and Its Applications 109,
DOI 10.1007/978-3-319-30785-5_4

variational inequality, without any loss of precision for the purposes of the second-order shape-topological sensitivity analysis. The second-order shape-topological directional derivatives are obtained for the Laplacian and for linear elasticity in two and three spatial dimensions. In the proposed method of sensitivity analysis, the singular geometrical perturbations $\epsilon \to \omega_\epsilon \subset \Omega$ centred at $\hat{x} \in \Omega$ are replaced by regular perturbations of bilinear forms supported on the manifold $\Gamma_R = \{|x - \hat{x}| = R\}$ in an elastic body, with $R > \epsilon > 0$. The obtained expressions for topological derivatives are easy to compute and therefore useful in numerical methods of topological optimization for contact problems.

1 Introduction

Topological derivatives of shape functionals $\Omega \to J(\Omega)$ are introduced in [25] for linear elliptic boundary value problems [6] defined in singularly perturbed domains $\epsilon \to \Omega(\epsilon)$, where $\epsilon \to 0$ is a small parameter which governs the size of small hole or inclusion in the bounded domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$. The topological derivatives are given by expressions depending on pointwise values of solutions as well as of its gradients [22]. Therefore, the obtained expressions for topological derivatives are not well defined in the energy spaces associated with the boundary value problems under considerations.

In this paper the topological sensitivity analysis of solutions to variational inequalities is performed by a domain decomposition technique. The regular perturbations defined on the energy space $\varepsilon \to \varepsilon^d b(\Gamma_R; \cdot, \cdot)$ for bilinear forms $\varepsilon \to a(\Omega(\epsilon); \cdot, \cdot)$ of boundary value problems, with respect to small parameter $\epsilon \to 0$, are introduced. Such perturbations are given by line integrals in two spatial dimensions, or by surface integrals in three spatial dimensions. As a result, the topological derivatives of shape functionals can be derived for solutions of variational inequalities posed in the intact domain Ω.

In order to derive the topological derivatives by an application of the domain decomposition technique the artificial interface $\Sigma \subset \Omega$ is introduced and $\Omega := \Omega_1 \cup \Sigma \cup \Omega_2$ is decomposed into two subdomains.

For the boundary value problem under considerations such a decomposition is indeed useful. In some applied problems we are interested in the influence of singular perturbations in subdomain Ω_1 on the behaviour of solutions in subdomain Ω_2. The functional under consideration is the elastic energy $\mathcal{E}(\Omega)$ of whole domain Ω. The mixed second-order derivatives of shape-topological or topological-shape types for the elastic energy are evaluated. The shape sensitivity analysis is performed e.g., in Ω_2, then the asymptotic analysis is performed in the second subdomain Ω_1. In the framework of shape-topological sensitivity analysis the velocity method is applied in order to determine the shape functional $J(\Omega) := d\mathcal{E}(\Omega; V)$, where V is the specific vector field in derivation of $V \to d\mathcal{E}(\Omega; V)$. Then the asymptotic expansion of $\epsilon \to J(\Omega_\epsilon)$ is evaluated. In the framework of topological-shape sensitivity analysis, first the asymptotic expansion of $\epsilon \to \mathcal{E}(\Omega_\epsilon)$ is performed, and the first order term

of such an expansion is called the topological derivative. It turns out [22, 25] that the topological derivative of energy functional is not well defined for arbitrary elements from the energy space of the elasticity boundary value problems under considerations. Therefore, we introduce the equivalent representations of topological derivatives which are well defined in the energy space. These representations can be used as well to modify the state equations by replacing the singular domain perturbations by the regular perturbations of bilinear forms in variational setting.

The asymptotic expansion of the energy functional performed in one subdomain, e.g., Ω_1, can be used in the second subdomain Ω_2 to evaluate the asymptotic expansion of the Steklov–Poincaré operator on the interface between subdomains. The method is justified by the fact that the first order expansion of the energy functional in the subdomain leads to the first order asymptotic expansion of the Dirichlet-to-Neumann mapping on the interface between subdomains. Thus, the first order expansion of the Steklov–Poincaré operator on the interface for the second subdomain is obtained. In this way the first order expansion of the energy functional in the truncated domain Ω_2 is derived. The precision of the obtained expansion is sufficient [27, 28] to replace the original energy functional by its first order expansion, provided the obtained expression is well defined on the energy space. Furthermore, the first order approximation of the energy functional in Ω is established. We point out that another method of approximation of the state equation by using the so-called self-adjoint extensions of the elliptic operators can be considered [20, 21].

1.1 Asymptotic Approximation for Variational Inequalities

The proposed domain decomposition method is important for variational inequalities. The asymptotic analysis of solutions to variational inequalities is more involved [3] compared to the analysis of solutions to linear elliptic boundary value problems.

The variational inequality under consideration results from the minimization problem of quadratic functional

$$v \to I(v) = \frac{1}{2}a(v,v) - L(v) \tag{1}$$

over a convex, closed subset $K \subset H$ of the Hilbert space H called the energy space. The function space $H := H(\Omega)$ is a Sobolev space which contains the functions defined over a domain $\Omega \subset \mathbb{R}^d$, $d = 2,3$. The singular geometrical perturbation ω_ϵ centred at $\hat{x} \in \Omega$ of the domain Ω is denoted by Ω_ϵ, the size of perturbation is governed by a small parameter $\epsilon \to 0$. The simple example of such a perturbation is the hole or inclusion at the origin $B_\epsilon := \{|x| < \epsilon\}$.

The quadratic functional defined on $H := H(\Omega_\epsilon)$ becomes

$$v \to I_\epsilon(v) = \frac{1}{2} a_\epsilon(v,v) - L_\epsilon(v) \tag{2}$$

with the minimizers denoted by $u_\epsilon \in K := K(\Omega_\epsilon)$.

The expansion of associated energy functional

$$\epsilon \to \mathcal{E}(\Omega_\epsilon) := I_\epsilon(u_\epsilon) = \frac{1}{2} a_\epsilon(u_\epsilon, u_\epsilon) - L_\epsilon(u_\epsilon) \tag{3}$$

is considered at $\epsilon = 0$.

Namely, we are looking for its asymptotic expansion

$$\mathcal{E}(\Omega_\epsilon) = \mathcal{E}(\Omega) + \epsilon^d \mathcal{T}(\hat{x}) + o(\epsilon^d), \tag{4}$$

where $\hat{x} \to \mathcal{T}(\hat{x})$ is the topological derivative [22, 25]. We show that there are regular perturbations of the bilinear form defined on the energy space $H(\Omega)$,

$$v \to b(v,v)$$

such that the perturbed quadratic functional defined on the unperturbed function space $H(\Omega)$

$$v \to I^\epsilon(v) = \frac{1}{2} \left[a(v,v) + \epsilon^d b(v,v) \right] - L(v) \tag{5}$$

furnishes the first order expansion (4). In our applications to contact problems in linear elasticity it turns out that the bilinear form $v \to b(v,v)$ is supported on $\Gamma_R := \{|x - \hat{x}| = R\} \subset \Omega$ with $R > \epsilon > 0$.

Remark 1. The contact problems in elastic bodies are modeled by variational inequalities

$$u \in K \ : \ a(u, v-u) \geq L(v-u) \ \forall v \in K. \tag{6}$$

For the sensitivity analysis in singularly perturbed geometrical domains, the weak solutions of contact problems $\epsilon \to u_\epsilon$ are given by perturbed variational inequalities

$$u \in K \ : \ a(u, v-u) + \epsilon^d b(u, v-u) \geq L(v-u) \ \forall v \in K, \tag{7}$$

where $\epsilon \to 0$ measures the size of singular perturbation. This is the main contribution of the paper. Therefore, we need the form of $\epsilon^d b(u, v-u)$ in order to apply our method of sensitivity analysis to numerical methods of topological optimization.

Variational inequalities are used to model contact problems in elasticity. It is known that the solutions to variational inequalities are Lipschitz continuous with respect to the shape [29]. In general, the state governed by a variational

inequality is not Fréchet differentiable with respect to the shape. For a class of variational inequalities described by the unilateral constraints in Sobolev spaces of Dirichlet type the metric projection onto the constraints turns out to be Hadamard differentiable [7]. This property is used in order to obtain the first order directional differentiability of the associated shape functionals.

In order to show the second-order shape differentiability for variational inequalities, we have to restrict ourselves to energy-type shape functionals. The energy functional is the so-called marginal function and it is Fréchet differentiable with respect to the shape [7]. The first order shape derivative of the energy functional in the direction of a specific velocity vector field is considered as the shape functional for topological optimization. Thus, its topological derivative is evaluated.

The possible applications of shape-topological derivatives include the control of singularities of solutions to variational inequalities by insertion of elastic inclusions far from the singularities.

We describe the shape-topological differentiability of the energy shape functional for the Signorini problem in two spatial dimensions. The same idea can be used for the frictionless contact problems in linear elasticity.

Let us consider the Signorini problem posed in $\Omega \subset \mathbb{R}^2$, with boundary $\partial\Omega = \Gamma \cup \Gamma_0$, and $\Gamma_c \subset \Gamma$. Denote $H^1_{\Gamma_0}(\Omega) = \{v \in H^1(\Omega) \mid v = 0 \text{ on } \Gamma_0 \subset \partial\Omega\}$.

The solution $u \in K$ minimizes the quadratic functional

$$I(v) = \frac{1}{2} a(\Omega; v, v) - (f, v)_\Omega$$

over the cone

$$K = \{v \in H^1_{\Gamma_0}(\Omega) \mid v \geq 0 \text{ on } \Gamma_c \subset \Gamma \subset \partial\Omega\}.$$

The shape functional is the energy

$$\mathcal{E}(\Omega) = \frac{1}{2} a(\Omega; u, u) - (f, u)_\Omega,$$

where

$$a(\Omega; u, u) = \int_\Omega \nabla u \cdot \nabla u dx,$$

$$(f, u)_\Omega = \int_\Omega f u dx.$$

We assume that $\overline{\Gamma} \cap \overline{\Gamma}_0 = \emptyset$. Let $\Gamma_0^t := T_t(V)(\Gamma_0)$ be the boundary variations [29] of the Dirichlet boundary Γ_0.

Let us consider the decomposition of $\Omega = \Omega_1 \cup \Sigma \cup \Omega_2$, $\overline{\Omega}_1 \cap \overline{\Omega}_2 = \emptyset$, such that $\Gamma_0 \subset \partial\Omega_1$ and $\Gamma_c \subset \partial\Omega_2$. It means that the boundary variations as well as the topological asymptotic analysis are performed in Ω_1, and the unilateral conditions are prescribed in the second subdomain Ω_2.

The shape derivative of the energy functional with respect to the boundary variations of Γ_0 can be written in distributed form [29]

$$d\mathcal{E}(\Omega;V) = \int_{\Omega_1} \langle A'(0)\cdot\nabla u, \nabla u\rangle dx$$

where $A'(0) = \operatorname{div} VI - DV - DV^\top$, under the assumption that the velocity field V is supported in a small neighbourhood of Γ_0 and that supp $V \cap$ supp f $= \emptyset$.

The second shape functional for the purposes of topological optimization is simply defined by

$$J(\Omega) := \int_{\Omega_1} \langle A'(0)\cdot\nabla u, \nabla u\rangle dx. \tag{8}$$

We are going to determine the topological derivatives of $\Omega \to J(\Omega)$ for insertion of small inclusions in Ω_1 far from Γ_0. In this way we could control the possible singularities on Γ_0 by topology optimization in Ω.

We consider the domain decomposition method for purposes of the shape-topological differentiability of energy shape functionals. First, the domain Ω is split into two subdomains Ω_1, Ω_2 and the interface Σ. The differentiability with respect to small parameter of the Dirichlet-to-Neumann map which lives on the boundary $\Sigma \subset \partial\Omega_1$ is established. This map is called the Steklov–Poincaré operator for subdomain Ω_2.

Once, the derivative of the energy functional is given, we can proceed with the subsequent topological optimization problem. For topological optimization another decomposition $\Omega := \Omega_R \cup \Gamma_R \cup \Omega_c$ is introduced. The small inclusion ω_ε centred at the origin $\hat{x} := \mathcal{O}$ is located in subdomain $\Omega_R \subset \Omega$ with the interface $\Gamma_R \subset \partial\Omega_R$.

2 Applications of Steklov–Poincaré Operators in Asymptotic Analysis

We analyse the precision of the proposed method of approximation for variational inequalities in singularly perturbed geometrical domains. We assume for simplicity that the singular perturbation is a disc $B_\epsilon = \{|x| < \epsilon\}$.

The Signorini variational inequality in $\Omega_\epsilon := \Omega \setminus \overline{B}_\epsilon$,

$$u_\epsilon \in K(\Omega_\epsilon) \ : \ a(\Omega_\epsilon; u_\epsilon, v - u_\epsilon) - L(\Omega_\epsilon; v - u_\epsilon) \geq 0 \quad \forall v \in u_\epsilon \in K(\Omega_\epsilon), \tag{9}$$

can be considered in the truncated domain $\Omega_c := \Omega \setminus \overline{B}_R$ for $R > \epsilon > 0$, R small enough. It is assumed that the source or linear form $v \to L(\Omega; v) := (f, v)_{\Omega_c}$ is supported in Ω_c. Hence the restriction $u_\epsilon \in K(\Omega_c)$ of $u_\epsilon \in K(\Omega_\epsilon)$ to the truncated domain is given by the solution to variational inequality

$$u_\epsilon \in K(\Omega_c) \ : \ a(\Omega_c; u_\epsilon, v-u_\epsilon) + \langle \mathcal{A}_\epsilon(u_\epsilon), v-u_\epsilon\rangle - L(\Omega_c; v-u_\epsilon) \geq 0 \quad \forall v \in K(\Omega_\epsilon), \tag{10}$$

where $\mathcal{A}_\epsilon$ stands for the Steklov–Poincaré operator which *replaces* the portion of bilinear form defined over the ring $C(R,\epsilon) := \{R > |x| > \epsilon\}$.

Proposition 1. *Assume that the Steklov–Poincaré operator admits the one-term expansion*

$$\langle \mathcal{A}_\epsilon(v), v\rangle = \langle \mathcal{A}(v), v\rangle + \epsilon^2 \langle \mathcal{B}(v), v\rangle + o(\epsilon^2; v, v) \tag{11}$$

with the compact remainder $o(\epsilon^2; v, v)$, then we can replace in (10) the Steklov–Poincaré operator by its one term approximation

$$\begin{aligned} \tilde{u}_\epsilon \in K(\Omega_c) \ : \ & a(\Omega_c; \tilde{u}_\epsilon, v-\tilde{u}_\epsilon) + \langle \mathcal{A}(\tilde{u}_\epsilon), v-\tilde{u}_\epsilon\rangle \\ & + \epsilon^2 \langle \mathcal{B}(\tilde{u}_\epsilon), v-\tilde{u}_\epsilon\rangle - L(\Omega_c; v-\tilde{u}_\epsilon) \geq 0 \quad \forall v \in K(\Omega_\epsilon), \end{aligned} \tag{12}$$

with the estimate

$$\|\tilde{u}_\epsilon - u_\epsilon\| = o(\epsilon^2). \tag{13}$$

Remark 2. From Proposition 1 it follows that for the shape-topological differentiability of the energy functional we can consider the variational inequality

$$\hat{u}_\epsilon \in K(\Omega) \ : \ a(\Omega; \hat{u}_\epsilon, v-\hat{u}_\epsilon) + \epsilon^2 \langle \mathcal{B}(\hat{u}_\epsilon), v-\hat{u}_\epsilon\rangle - L(\Omega; v-\hat{u}_\epsilon) \geq 0 \quad \forall v \in K(\Omega), \tag{14}$$

since $\|\hat{u}_\epsilon - u_\epsilon\| = o(\epsilon^2)$ in Ω_c.

In this way, the approximation (5) of quadratic functional (2) is justified for the first order topological derivatives of variational inequalities in truncated domains.

For the quadratic functional (1) and the associated boundary value problem, the bilinear form

$$v \to b(\Gamma_R; v, v) := \langle \mathcal{B}(v), v\rangle$$

is determined. The linear operator $\mathcal{B}$ is obtained from the one term expansion of the Steklov–Poincaré operator $\mathcal{A}_\epsilon$, the expansion results from the energy expansion in the subdomain Ω_R. Therefore, the perturbed quadratic functional (3) can be replaced by its approximation given by (5). For the Signorini problem in two spatial dimensions it means that the variational inequality is obtained for minimization of the perturbed functional (3) over the energy space in unperturbed domain Ω, and the associated energy functional

$$\mathcal{E}_\epsilon(\Omega) = \frac{1}{2} a(\Omega; u_\epsilon, u_\epsilon) + \frac{\epsilon^2}{2} b(\Gamma_R; u_\epsilon, u_\epsilon) - (f, u_\epsilon)_\Omega,$$

is evaluated for the solution of variational inequality

$$u_\epsilon \in K(\Omega) \ : \ a(\Omega; u_\epsilon, v-u_\epsilon) + \epsilon^2 b(\Gamma_R; u_\epsilon, v-u_\epsilon) - (f, v-u_\epsilon)_\Omega \geq 0 \quad \forall v \in K(\Omega).$$

3 Asymptotic Analysis by Domain Decomposition Method

In order to apply the domain decomposition technique to topological differentiability $\omega_\epsilon \to J_\epsilon(\Omega)$ in topologically perturbed domains $\Omega := \Omega_\epsilon$ for the shape functionals $\Omega \to J(\Omega)$, we need the appropriate results on topological differentiability $\epsilon \to \mathcal{B}_\epsilon$ of the Steklov–Poincaré pseudodifferential boundary operators defined on the artificial interface Σ. In the particular case of holes $\epsilon \to \omega_\epsilon$ the notation is straightforward, with the singularly perturbed domain $\Omega_\epsilon := \Omega \setminus \overline{\omega}_\epsilon$ and with the shape functional to be analysed with respect to small parameter $\epsilon \to J_\epsilon(\Omega) := J(\Omega \setminus \overline{\omega}_\epsilon)$. In the case of inclusions $\epsilon \to \omega_\epsilon$ the shape functional depends on the characteristic functions $\epsilon \to \chi_\epsilon$ of the domain perturbation ω_ϵ. For inclusions the state solution $\epsilon \to u_\epsilon \in H(\Omega)$ is obtained by solving boundary value problems with operator coefficients depending on the small parameter $\epsilon \to 0$. In both cases the asymptotics of Steklov–Poincaré operators are obtained by asymptotic analysis of the energy functional for linear elliptic boundary value problems in subdomains Ω_2 which contains the perturbations $\epsilon \to \omega_\epsilon$.

Let us consider the direct method of sensitivity analysis in subdomain Ω_1 which contains the contact subset $\Gamma_c \subset \partial\Omega$. This is possible due to the conical differentiability of metric projection onto the convex set K which is valid under some assumptions (e.g., the convex, closed cone K is polyhedric in the Dirichlet space $H(\Omega)$ [7]).

In the case of the Signorini problem in two spatial dimensions the direct method of asymptotic analysis for the shape functional (8)

$$J_\epsilon(\Omega_\epsilon) := \int_{\Omega_1} \langle A'(0) \cdot u_\epsilon, u_\epsilon \rangle dx$$

can be described as follows for the disc $\omega_\epsilon := B(\epsilon) = \{|x| < \epsilon\}$ located at the origin.

1. We solve the variational inequality in Ω_1 : determine $u \in K$ and its coincidence set $\Xi := \{x \in \Gamma_c \ : \ u(x) = 0\}$. Thus, the convex cone

$$S = \{v \in H^1_{\Gamma_0}(\Omega) \ : \ v \geq 0 \text{ on } \Xi, \quad a(\Omega; u, v) = (f, v)_\Omega\}$$

   used in conical differentiability of the element u with respect to the shape can be determined.
2. The asymptotic analysis of solutions to variational inequality in singularly perturbed domain $\Omega(\epsilon) : \Omega \setminus \overline{B(\epsilon)}$ with respect to small parameter $\epsilon \to 0$ which governs the size of the hole $B(\epsilon)$ leads to the expansion

$$u_\epsilon = u + \epsilon^2 q + o(\epsilon^2)$$

obtained by the domain decomposition method with the Steklov–Poincaré boundary operators, where

$$q \in S \ : \ a(\Omega; q, v-q) + \epsilon^2 \langle \mathcal{B}q, v-q \rangle_R \geq \quad \forall v \in S.$$

3. The shape functional

$$J_\epsilon(\Omega_\epsilon) := \int_{\Omega_1} \langle A'(0) \cdot u_\epsilon, u_\epsilon \rangle dx$$

can be expanded in Ω_1, the expansion is valid in the whole domain Ω,

$$J_\epsilon(\Omega_\epsilon) = \int_\Omega \langle A'(0) \cdot u, u \rangle dx + 2\epsilon^2 \int_\Omega \langle A'(0) \cdot q, u \rangle dx + o(\epsilon^2),$$

however the obtained expression for the topological derivative may not be constructive in numerical methods. We want to obtain an equivalent expression, when possible, which replaces the topological derivative

$$\mathcal{T}(\mathcal{O}) = 2 \int_\Omega \langle A'(0) \cdot q, u \rangle dx$$

in the first order expansion of the energy functional for Signorini problem. In the linear boundary value problems such an expression can always be obtained by the introduction of an appropriate adjoint state. We point out that for variational inequalities the existence of an adjoint state cannot be expected in general.

4 Asymptotic Analysis of Boundary Value Problems in Rings or Spherical Shells

4.1 Elasticity Boundary Value Problems

In this section we shall consider asymptotic corrections to the energy function corresponding to the elasticity system or Laplace equation in $\mathbb{R}^d$, where $d = 2, 3$. The change of the energy is caused by creating a small ball-like void of variable radius ϵ in the interior of the domain Ω, with homogeneous Neumann boundary condition on its surface. We assume that this void has its centre at the origin $\mathcal{O}$. In order to eliminate the variability of the domain, we take as Ω_R the open ball $B(\mathcal{O}, R) = B(R)$ with fixed R. In this way the void $B(\epsilon)$ is surrounded by $B(R) \subset \operatorname{int}\Omega$. We denote also the ring or spherical shell as $C(R, \epsilon) = B(R) \setminus \overline{B}(\epsilon)$, $\Omega(R) = \Omega \setminus \overline{B}(R)$ and $\Gamma_R = \partial B(R)$.

Using these notations we define our main tool, namely the Dirichlet-to-Neumann mapping for linear elasticity or the Steklov–Poincaré operator

$$\mathcal{A}_\epsilon : \mathbf{H}^{1/2}(\Gamma_R) \longmapsto \mathbf{H}^{-1/2}(\Gamma_R)$$

by means of the boundary value problem:

$$\begin{aligned} &(1-2\nu)\Delta\mathbf{w} + \mathbf{grad}\,\mathrm{div}\,\mathbf{w} = 0, \ \text{ in } \ C(R,\epsilon), \\ &\mathbf{w} = \mathbf{v} \ \text{ on } \ \Gamma_R, \\ &\sigma(\mathbf{w}).\mathbf{n} = 0 \ \text{ on } \ \Gamma_\epsilon \end{aligned} \tag{15}$$

so that

$$\mathcal{A}_\epsilon \mathbf{v} = \sigma(\mathbf{w}).\mathbf{n} \ \text{ on } \ \Gamma_R. \tag{16}$$

Domain Decomposition: Steklov–Poincaré Operator Let $\mathbf{u}^R$ be the restriction of $\mathbf{u}$ to $\Omega(R)$ and $\gamma^R\varphi$ the projection of φ on Γ_R. We may then define the functional

$$\begin{aligned} I^R_\epsilon(\varphi_\epsilon) = \frac{1}{2}\int_{\Omega(R)} \sigma(\varphi_\epsilon) : \varepsilon(\varphi_\epsilon)\,dx - \int_{\Gamma_N} \mathbf{h}.\varphi_\epsilon\,ds + \\ + \frac{1}{2}\int_{\Gamma_R} (\mathcal{A}_\epsilon \gamma^R \varphi_\epsilon).\gamma^R\varphi_\epsilon\,ds \end{aligned} \tag{17}$$

and the solution $\mathbf{u}^R_\epsilon$ as a minimal argument for

$$I^R_\epsilon(\mathbf{u}^R_\epsilon) = \inf_{\varphi_\epsilon \in K \subset V_\epsilon} I^R_\epsilon(\varphi_\epsilon), \tag{18}$$

Here lies the essence of the domain decomposition concept: we have replaced the variable domain by a fixed one, at the price of introducing variable boundary operator $\mathcal{A}_\epsilon$.

The above expressions have even simpler form in case of a single Laplace equation. It is enough to replace the displacement by the scalar function u, elasticity operator by $-\Delta$, and

$$\sigma(\mathbf{u}) := \mathbf{grad}\,u, \qquad \varepsilon(\mathbf{u}) := \mathbf{grad}\,u, \qquad \sigma(\mathbf{u}).\mathbf{n} := \partial u/\partial\mathbf{n}.$$

The goal is to find the expansion

$$\mathcal{A}_\epsilon = \mathcal{A} + \epsilon^d\mathcal{B} + \mathcal{R}_\epsilon, \tag{19}$$

where the remainder $\mathcal{R}_\epsilon$ is of order $o(\epsilon^d)$ in the operator norm in the space $L(\mathbf{H}^{1/2}(\Gamma_R), \mathbf{H}^{-1/2}(\Gamma_R))$, and the operator $\mathcal{B}$ is regular enough, namely it is bounded and linear:

$$\mathcal{B} \in L\big(\mathbf{L}_2(\Gamma_R), \mathbf{L}_2(\Gamma_R)\big).$$

Under this assumption the following propositions hold.

Proposition 2. *Assume that* (19) *holds in the operator norm. Then strong convergence takes place*

$$\mathbf{u}_\epsilon^R \to \mathbf{u}^R \tag{20}$$

in the norm of $\mathbf{H}^1(\Omega(R))$.

Proposition 3. *The energy functional has the representation*

$$I_\epsilon^R(\mathbf{u}_\epsilon^R) = I^R(\mathbf{u}^R) + \epsilon^d \langle \mathcal{B}(\mathbf{u}^R), \mathbf{u}^R \rangle_R + o(\epsilon^3)\ , \tag{21}$$

where $o(\epsilon^d)/\epsilon^d \to 0$ *with* $\epsilon \to 0$ *in the same energy norm.*

Here $I^R(\mathbf{u}^R)$ denotes the functional I_ϵ^R on the intact domain, i.e. $\epsilon := 0$ and $\mathcal{A}_\epsilon := \mathcal{A}$, applied to truncation of $\mathbf{u}$.

Generally, the energy correction for both elasticity system and Laplace operator has the form

$$\langle \mathcal{B}(\mathbf{u}^R), \mathbf{u}^R \rangle_R \ = \ -c_d e_u(\mathcal{O}),$$

where $c_d = \mathrm{vol}(B(1))$ with $B(1)$ being the unit ball in $\mathbb{R}^d$. The energy-like density function $e_u(\mathcal{O})$ has the form:

- In case of the Laplace operator

$$e_u(\mathcal{O}) \ = \ \frac{1}{2}\|\nabla u^R(\mathcal{O})\|^2$$

  for both $d = 2$ and $d = 3$, see [27].
- In case of the elasticity system

$$e_u(\mathcal{O}) \ = \ \frac{1}{2}\mathbb{P}\sigma(\mathbf{u}^R(\mathcal{O}) : \varepsilon(\mathbf{u}^R(\mathcal{O}),$$

  where for $d = 2$ and plain stress

$$\mathbb{P} \ = \ \frac{1}{1-\nu}(4\mathbb{I} - \mathrm{I} \otimes \mathrm{I})$$

  and for $d = 3$

$$\mathbb{P} \ = \ \frac{1-\nu}{7-5\nu}(10\mathbb{I} - \frac{1-5\nu}{1-2\nu}\mathrm{I} \otimes \mathrm{I})$$

see [22, 26]. Here $\mathbb{I}$ is the fourth order identity tensor, and I is the second-order identity tensor.

This approach is important for variational inequalities since it allows us to derive the formulas for topological derivatives which are similar to the expressions obtained for the corresponding linear boundary value problems.

4.2 Explicit form of the Operator $\mathcal{B}$ for the Laplacian in Two Spatial Dimensions

If the function u is harmonic in a ball $B(R) \subset \mathbb{R}^2$, of radius $R > 0$ and centre at $\mathbf{x}_0 = \mathcal{O}$, then the exact expressions for the first order derivatives of u take on the following form [27]

$$u_{/1}(\mathcal{O}) = \frac{1}{\pi R^3} \int_{\Gamma_R} u \cdot x_1 \, ds,$$
$$u_{/2}(\mathcal{O}) = \frac{1}{\pi R^3} \int_{\Gamma_R} u \cdot x_2 \, ds.$$

Since the line integrals on Γ_R are well defined for functions in $L_2(\Gamma_R)$, it follows that the operator $\mathcal{B}$ can be extended to the bounded operator on $L_2(\Gamma_R)$,

$$\mathcal{B} \in \mathcal{L}(L_2(\Gamma_R) \to L_2(\Gamma_R)).$$

The symmetric bilinear form for this operator, given by

$$\langle \mathcal{B}u, v \rangle_R =$$
$$-\frac{1}{2\pi R^6} \left[\left(\int_{\Gamma_R} u x_1 \, ds \right) \left(\int_{\Gamma_R} v x_1 \, ds \right) + \left(\int_{\Gamma_R} u x_2 \, ds \right) \left(\int_{\Gamma_R} v x_2 \, ds \right) \right]$$

is continuous for all $u, v \in L_2(\Gamma_R)$. In fact, the bilinear form

$$L_2(\Gamma_R) \times L_2(\Gamma_R) \ni (u, v) \mapsto b(\Gamma_R; u, v) \in \mathbb{R}$$

is continuous with respect to the weak convergence because of the simple structure

$$b(\Gamma_R; u, v) = l_1(u) l_1(v) + l_2(u) l_2(v) \quad u, v \in L_1(\Gamma_R)$$

with two linear forms $v \to l_i(v), i = 1, 2,$

$$l_i(u) = \frac{1}{\sqrt{2\pi}} R^{-3} \int_{\Gamma_R} u x_i \, ds$$

defined as line integrals on Γ_R. This gives an additional regularity for the regular nonlocal perturbation $\mathcal{B}$ of the pseudo-differential Steklov–Poincaré boundary operator $\mathcal{A}_\epsilon$.

4.3 Explicit form of the Operator $\mathcal{B}$ for the Laplacian in Three Spatial Dimensions

Similarly as in two spatial dimensions, for harmonic functions in $\mathbb{R}^3$, it may be proved [27] that

$$u_{/1}(\mathcal{O}) = \frac{3}{4\pi R^4} \int_{S(R)} ux_1 \, ds,$$
$$u_{/2}(\mathcal{O}) = \frac{3}{4\pi R^4} \int_{S(R)} ux_2 \, ds,$$
$$u_{/3}(\mathcal{O}) = \frac{3}{4\pi R^4} \int_{S(R)} ux_3 \, ds.$$

Using this one can easily write down the bilinear form

$$b(\Gamma_R; u, v) = \langle \mathcal{B}u, v \rangle_R = l_1(u)l_1(v) + l_2(u)l_2(v) + l_3(u)l_3(v),$$

where

$$l_i(u, v) \; = \; \sqrt{\frac{3}{8\pi}} R^{-4} \int_{S(R)} ux_i \, ds.$$

From the computational point of view, the effort in comparison with the two-dimensional case grows similarly as the difficulty of computing integrals over circle versus integrals over sphere.

4.4 Explicit form of the Operator $\mathcal{B}$ for Elasticity in Two Spatial Dimensions

Let us denote for the plain stress case

$$k = \frac{\lambda + \mu}{\lambda + 3\mu}.$$

It has been proved in [27] that the following exact formulae hold

$$\varepsilon_{11}(\mathcal{O})+\varepsilon_{22}(\mathcal{O}) = \frac{1}{\pi R^3}\int_{\Gamma_R}(u_1x_1+u_2x_2)\,ds,$$
$$\varepsilon_{11}(\mathcal{O})-\varepsilon_{22}(\mathcal{O}) = \frac{1}{\pi R^3}\int_{\Gamma_R}\left[(1-9k)(u_1x_1-u_2x_2)+\frac{12k}{R^2}(u_1x_1^3-u_2x_2^3)\right]ds,$$
$$2\varepsilon_{12}(\mathcal{O}) = \frac{1}{\pi R^3}\int_{\Gamma_R}\left[(1+9k)(u_1x_2+u_2x_1)-\frac{12k}{R^2}(u_1x_2^3+u_2x_1^3)\right]ds.$$

These expressions are easy to compute numerically, but contain additional integrals of third powers of x_i. Therefore, strains $\varepsilon_{ij}(\mathcal{O})$ may be expressed as linear combinations of integrals over circle which have the form

$$\int_{\Gamma_R}u_ix_j\,ds, \qquad \int_{\Gamma_R}u_ix_j^3\,ds.$$

The same is true, due to Hooke's law, for stresses $\sigma_{ij}(\mathcal{O})$. They may then be substituted into expression for the operator B, yielding

$$\langle\mathcal{B}(\mathbf{u}^R),\mathbf{v}^R\rangle_R = -\frac{1}{2}c_2\mathbb{P}\sigma(\mathbf{u}):\varepsilon(\mathbf{v}).$$

These formulas are quite similar to the ones obtained for Laplace operator and easy to compute numerically.

4.5 Explicit form of the Operator $\mathcal{B}$ for Elasticity in Three Spatial Dimensions

It turns out that similar situation holds in three spatial dimensions, but obtaining the formulas is more difficult. Assuming given values of $\mathbf{u}$ on Γ_R, the solution of elasticity system in $B(R)$ may be expressed, following partially the derivation from [17] (pages 285 and later), as

$$\mathbf{u} = \sum_{n=0}^{\infty}[\mathbf{U}_n+(R^2-r^2)k_n(\nu)\mathbf{grad}\,\mathrm{div}\,\mathbf{U}_n], \tag{22}$$

where $k_n(\nu) = 1/2[(3-2\nu)n-2(1-\nu)]$ and $r=\|\mathbf{x}\|$. In addition

$$\mathbf{U}_n = \frac{1}{R^n}[\mathbf{a}_{n0}d_n(\mathbf{x})+\sum_{m=1}^{n}(\mathbf{a}_{nm}c_n^m(\mathbf{x})+\mathbf{b}_{nm}s_n^m(\mathbf{x}))]. \tag{23}$$

The vectors

$$\mathbf{a}_{n0} = (a_{n0}^1, a_{n0}^2, a_{n0}^3)^\top,$$

$$\mathbf{a}_{nm} = (a^1_{nm}, a^2_{nm}, a^3_{nm})^\top,$$
$$\mathbf{b}_{nm} = (b^1_{nm}, b^2_{nm}, b^3_{nm})^\top$$

are constant and the set of functions

$$\{d_0;\ d_1, c^1_1, s^1_1;\ d_2, c^1_2, s^1_2, c^2_2, s^2_2;\ d_3, c^1_3, s^1_3, c^2_3, s^2_3, c^3_3, s^3_3;\ \ldots\}$$

constitutes the complete system of orthonormal harmonic polynomials on Γ_R, related to Laplace spherical functions, see the next paragraph. Specifically,

$$c^l_k(\mathbf{x}) = \frac{\hat{P}^{l,c}_k(\mathbf{x})}{\|\hat{P}^{l,c}_k\|_R}, \quad s^l_k(\mathbf{x}) = \frac{\hat{P}^{l,s}_k(\mathbf{x})}{\|\hat{P}^{l,s}_k\|_R}, \quad d_k = \frac{P_k(\mathbf{x})}{\|\hat{P}_k\|_R}.$$

For example,

$$c^2_3(\mathbf{x}) = \frac{1}{R^4}\sqrt{\frac{7}{240\pi}}(15x^2_1x_3 - 15x^2_2x_3),$$

If the value of $\mathbf{u}$ on Γ_R is assumed as given, then, denoting

$$\langle \phi, \psi \rangle_R = \int_{\Gamma_R} \phi\,\psi\, ds,$$

we have for $n \geq 0$, $m = 1, \ldots, n$, $i = 1, 2, 3$:

$$\begin{aligned} a^i_{n0} &= R^n \langle u_i, d_n(\mathbf{x}) \rangle_R, \\ a^i_{nm} &= R^n \langle u_i, c^m_n(\mathbf{x}) \rangle_R, \\ b^i_{nm} &= R^n \langle u_i, s^m_n(\mathbf{x}) \rangle_R. \end{aligned} \tag{24}$$

Since we are looking for $\varepsilon_{ij}(\mathcal{O})$, only the part of $\mathbf{u}$ which is linear in $\mathbf{x}$ is relevant. It contains two terms:

$$\hat{\mathbf{u}} = \mathbf{U}_1 + R^2 k_3(\nu)\,\mathbf{grad}\,\mathrm{div}\,\mathbf{U}_3. \tag{25}$$

For any $f(\mathbf{x})$, $\mathbf{grad}\,\mathrm{div}\,(\mathbf{a}f) = H(f)\cdot\mathbf{a}$, where $\mathbf{a}$ is a constant vector and $H(f)$ is the Hessian matrix of f. Therefore

$$\begin{aligned} \hat{\mathbf{u}} = {} & \frac{1}{R}[\mathbf{a}_{10} d_1(\mathbf{x}) + \mathbf{a}_{11} c^1_1(\mathbf{x}) + \mathbf{b}_{11} s^1_1(\mathbf{x})] \\ & + R^2 k_3(\nu)\frac{1}{R^3}\Big[H(d_3)(\mathbf{x})\mathbf{a}_{30} \\ & + \sum_{m=1}^{3}\big(H(c^m_3)(\mathbf{x})\mathbf{a}_{3m} + H(s^m_3)(\mathbf{x})\mathbf{b}_{3m}\big)\Big] \end{aligned} \tag{26}$$

From the above we may single out the coefficients standing at x_1, x_2, x_3 in u_1, u_2, u_3. For example,

$$\varepsilon_{11}(\mathcal{O}) = \frac{1}{R^3}\sqrt{\frac{3}{4\pi}}a^1_{11} + \frac{1}{R^5}k_3(\nu)\Big[-3\sqrt{\frac{7}{4\pi}}a^3_{30} - 9\sqrt{\frac{7}{24\pi}}a^1_{31} - 3\sqrt{\frac{7}{24\pi}}b^2_{31} + 30\sqrt{\frac{7}{240\pi}}a^3_{32} + 90\sqrt{\frac{7}{1440\pi}}a^1_{33} + 90\sqrt{\frac{7}{1440\pi}}b^2_{33}\Big],$$

$$\varepsilon_{12}(\mathcal{O}) = \frac{1}{R^3}\sqrt{\frac{3}{4\pi}}(b^1_{11} + a^2_{11}) + \frac{1}{R^5}k_3(\nu)\Big[-3\sqrt{\frac{7}{24\pi}}a^2_{31} - \sqrt{\frac{7}{24\pi}}b^1_{31} + 15\sqrt{\frac{7}{60\pi}}b^3_{32} - 90\sqrt{\frac{7}{1440\pi}}a^2_{33} + 90\sqrt{\frac{7}{1440\pi}}b^1_{33}\Big].$$

Observe that

$$\varepsilon_{11}(\mathcal{O}) + \varepsilon_{22}(\mathcal{O}) + \varepsilon_{33}(\mathcal{O}) = \frac{1}{R^3}\sqrt{\frac{3}{4\pi}}\big(R\langle u_1, c^1_1\rangle_R + R\langle u_2, s^1_1\rangle_R + R\langle u_3, d_1\rangle_R\big)$$

and $c^1_1 = \frac{1}{R^2}\sqrt{\frac{3}{4\pi}}x_1$, $s^1_1 = \frac{1}{R^2}\sqrt{\frac{3}{4\pi}}x_2$, $d_1 = \frac{1}{R^2}\sqrt{\frac{3}{4\pi}}x_3$, exactly the same as for the case of Laplace equation. This should be expected, since $\operatorname{tr}\varepsilon$ is a harmonic function.

As a result, the operator $\mathbf{B}$ may be defined by the formula

$$\langle \mathcal{B}\mathbf{u}, \mathbf{u}\rangle_R = -c_3\mathbb{P}\sigma(\mathbf{u}(\mathcal{O})) : \varepsilon(\mathbf{u}(\mathcal{O}))$$

but the right-hand side consists of integrals of $\mathbf{u}$ multiplied by first and third order polynomials in x_i over Γ_R resulting from (24). This is a very similar situation as in two spatial dimensions. Thus, the new expressions for strains make it possible to rewrite $\mathcal{B}$ in the form possessing the desired regularity.

4.6 Laplace Spherical Polynomials

For $n = 1$:

$$\hat{P}_1(\mathbf{x}) = x_3, \quad \hat{P}^{1,c}_1(\mathbf{x}) = x_1, \quad \hat{P}^{1,s}_1(\mathbf{x}) = x_2,$$

$$\|\hat{P}_1\|_R = \|\hat{P}^{1,c}_1\|_R = \|\hat{P}^{1,s}_1\|_R = R^2\sqrt{\frac{4\pi}{3}},$$

and for $n = 3$:

$$\hat{P}_3(\mathbf{x}) = x_3^3 - \frac{3}{2}x_2^2x_3 - \frac{3}{2}x_1^2x_3, \qquad \|\hat{P}_3\|_R = R^4\sqrt{\frac{4\pi}{7}},$$

$$\hat{P}_3^{1,c}(\mathbf{x}) = 6x_1x_3^2 - \frac{3}{2}x_1^3 - \frac{3}{2}x_1x_2^2, \qquad \|\hat{P}_3^{1,c}\|_R = R^4\sqrt{\frac{24\pi}{7}},$$

$$\hat{P}_3^{1,s}(\mathbf{x}) = 6x_2x_3^2 - \frac{3}{2}x_2^3 - \frac{3}{2}x_1^2x_2, \qquad \|\hat{P}_3^{1,s}\|_R = R^4\sqrt{\frac{24\pi}{7}},$$

$$\hat{P}_3^{2,c}(\mathbf{x}) = 15x_1^2x_3 - 15x_2^2x_3, \qquad \|\hat{P}_3^{2,c}\|_R = R^4\sqrt{\frac{240\pi}{7}},$$

$$\hat{P}_3^{2,s}(\mathbf{x}) = 15x_1x_2x_3, \qquad \|\hat{P}_3^{2,s}\|_R = R^4\sqrt{\frac{60\pi}{7}},$$

$$\hat{P}_3^{3,c}(\mathbf{x}) = 15x_1^3 - 45x_1x_2^2, \qquad \|\hat{P}_3^{3,c}\|_R = R^4\sqrt{\frac{1440\pi}{7}},$$

$$\hat{P}_3^{3,s}(\mathbf{x}) = 45x_1^2x_2 - 15x_2^3, \qquad \|\hat{P}_3^{3,s}\|_R = R^4\sqrt{\frac{1440\pi}{7}},$$

5 Asymptotic Analysis of Steklov–Poincaré Operators in Reinforced Rings in Two Spatial Dimensions

In this section the similar asymptotic analysis of elliptic boundary value problems in subdomain $\Omega_R \in \mathbb{R}^2$ is performed, but we modify the situation, assuming that the hole is filled only partially, different material constituting a fixed part of it. In this way, we may consider double asymptotic transition, where both the size of the hole, and the proportion of the different material contained in it can vary. Mechanically, this situation corresponds e.g., to the hole with hardened walls.

The analysis is based again on exact representation of solutions and allows to obtain the perturbation of solutions, using the fact that these solutions may be considered as minimizers of energy functional. The method is also suitable for double asymptotic expansions of solutions as well as energy form. The ultimate goal is to use obtained formulas in the evaluation of topological derivatives for elliptic boundary value problems.

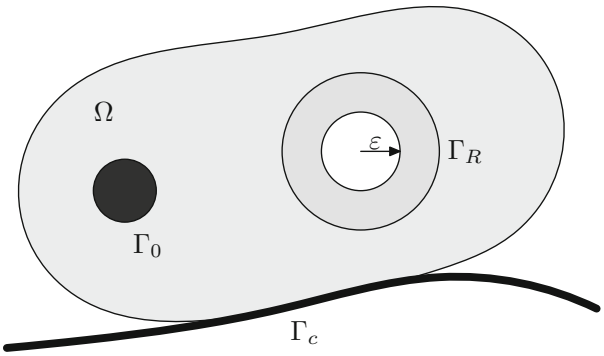

Fig. 1 The domain with the hole and the surrounding *circle*

5.1 Model Problem

Let us consider the domain Ω containing the hole with boundary made of modified material as depicted in Fig. 1. For simplicity the hole is located at the origin of the coordinate system. In order to write down the model problem, we introduce some notations.

$$\begin{aligned} B_s &= \{x \in \mathbb{R}^2 \,|\, \|x\| < s\} \\ C_{s,t} &= \{x \in \mathbb{R}^2 \,|\, s < \|x\| < t\} \\ \Gamma_s &= \{x \in \mathbb{R}^2 \,|\, \|x\| = s\} \\ \Omega_s &= \Omega \setminus B_s. \end{aligned}$$

Then the problem in the intact domain Ω has the form

$$\begin{aligned} k_1 \Delta w_0 &= 0 \quad \text{in } \Omega \\ w_0 &= g_0 \quad \text{on } \partial\Omega. \end{aligned} \tag{27}$$

The model problem in the modified domain reads:

$$\begin{aligned} k_1 \Delta w_\rho &= 0 \quad \text{in } \Omega_\rho \\ w_\rho &= g_0 \quad \text{on } \partial\Omega \\ w_\rho &= v_\rho \quad \text{on } \Gamma_\rho \\ k_2 \Delta v_\rho &= 0 \quad \text{in } C_{\lambda\rho,\rho} \end{aligned} \tag{28}$$

$$k_2 \frac{\partial v_\rho}{\partial n_2} = 0 \quad \text{on } \Gamma_{\lambda\rho}$$
$$k_1 \frac{\partial w_\rho}{\partial n_1} + k_2 \frac{\partial v_\rho}{\partial n_2} = 0 \quad \text{on } \Gamma_\rho,$$

where n_1—exterior normal vector to Ω_ρ, n_2—exterior normal vector to $C_{\lambda\rho,\rho}$, and $0 < \lambda < 1$.

We want to investigate the influence of the small ring-like inclusion made of another material on the difference $w_\rho - w_0$ in Ω_R, where Γ_R surrounds $C_{\lambda\rho,\rho}$ and R is fixed. We assume that $\rho \to 0+$ and λ is considered temporarily constant.

If we define

$$u_\rho = \begin{cases} w_\rho & \text{in } \Omega_\rho \\ v_\rho & \text{in } C_{\lambda\rho,\rho} \end{cases}$$

then the problem (28) reduces to finding the minimum of the energy functional

$$\mathcal{E}_1(u_\rho) = \frac{1}{2}\int_{\Omega_\rho} k_1 \nabla u_\rho \cdot \nabla u_\rho \, dx + \frac{1}{2}\int_{C_{\lambda\rho,\rho}} k_2 \nabla u_\rho \cdot \nabla u_\rho \, dx \tag{29}$$

for $u_\rho \in H^1(\Omega_\rho)$, $u_\rho = g_0$ on $\partial\Omega$.

This expression may be rewritten as

$$\begin{aligned} \mathcal{E}_1(u_\rho) = \frac{1}{2}\int_{\Omega_R} k_1 \nabla w_\rho \cdot \nabla w_\rho \, dx + \\ + \frac{1}{2}\int_{C_{\rho,R}} k_1 \nabla w_\rho \cdot \nabla w_\rho \, dx + \\ + \frac{1}{2}\int_{C_{\lambda\rho,\rho}} k_2 \nabla v_\rho \cdot \nabla v_\rho \, dx. \end{aligned}$$

Using integration by parts we obtain

$$\begin{aligned} \mathcal{E}_1(u_\rho) = \frac{1}{2}\int_{\Omega_R} k_1 \nabla w_\rho \cdot \nabla w_\rho \, dx + \\ + \frac{1}{2}\int_{\Gamma\rho} \left(w_\rho k_1 \frac{\partial w_\rho}{\partial n_1} + v_\rho k_2 \frac{\partial v_\rho}{\partial n_2} \right) ds + \\ + \frac{1}{2}\int_{\Gamma_R} k_1 w_\rho \frac{\partial w_\rho}{\partial n_3} \, ds, \end{aligned}$$

where n_3 is the exterior normal to Ω_R. Hence, due to boundary and transmission condition,

$$\mathcal{E}_1(u_\rho) = \frac{1}{2}\int_{\Omega_R} k_1 \nabla w_\rho \cdot \nabla w_\rho \, dx + \frac{1}{2}\int_{\Gamma_R} k_1 w_\rho \frac{\partial w_\rho}{\partial n_3}\, ds. \tag{30}$$

5.2 Steklov–Poincaré Operator

Observe that $\mathcal{E}_1(w_0)$ corresponds to the problem (27). Therefore the main goal is to find the Steklov–Poincaré operator

$$\mathcal{A}_{\lambda,\rho} \,:\, w \in H^{1/2}(\Gamma_R) \longmapsto \frac{\partial w_\rho}{\partial n_3} \in H^{-1/2}(\Gamma_R), \tag{31}$$

where the normal derivative is computed from auxiliary problem

$$\begin{aligned} k_1 \Delta w_\rho &= 0 \quad \text{in } C_{\rho,R} \\ w_\rho &= w \quad \text{on } \Gamma_R \\ w_\rho &= v_\rho \quad \text{on } \Gamma_\rho \\ k_2 \Delta v_\rho &= 0 \quad \text{in } C_{\lambda\rho,\rho} \\ k_2 \frac{\partial v_\rho}{\partial n_2} &= 0 \quad \text{on } \Gamma_{\lambda\rho} \\ k_1 \frac{\partial w_\rho}{\partial n_1} + k_2 \frac{\partial v_\rho}{\partial n_2} &= 0 \quad \text{on } \Gamma_\rho. \end{aligned} \tag{32}$$

The geometry of domains of definition for functions is shown in Fig. 2. Now let us adopt the polar coordinate system around origin and assume the Fourier series form for w on Γ_R.

$$w = C_0 + \sum_{k=1}^{\infty} (A_k \cos k\varphi + B_k \sin k\varphi) \tag{33}$$

The general form of the solution w_ρ is

$$w_\rho = A^w + B^w \log r + \sum_{k=1}^{\infty} (w_k^c(r)\cos k\varphi + w_k^s(r)\sin k\varphi), \tag{34}$$

where

$$w_k^c(r) = A_k^c r^k + B_k^c \frac{1}{r^k}, \quad w_k^s(r) = A_k^s r^k + B_k^s \frac{1}{r^k}.$$

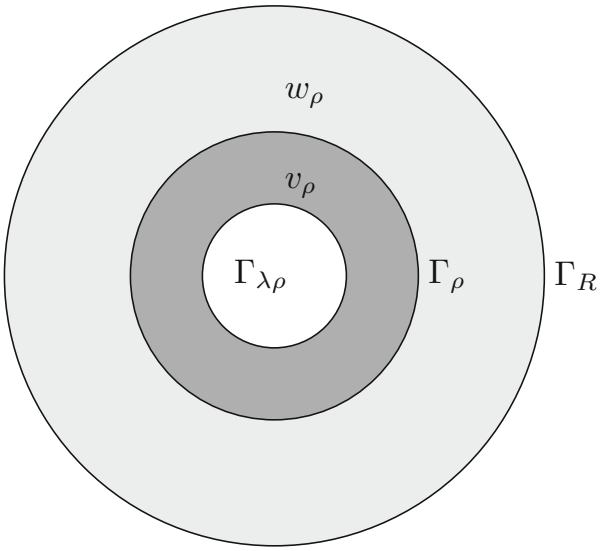

Fig. 2 Domains of definition for w_ρ and v_ρ

Similarly for v_ρ

$$v_\rho = A^v + B^v \log r + \sum_{k=1}^{\infty} \left(v_k^c(r)\cos k\varphi + v_k^s(r)\sin k\varphi\right), \tag{35}$$

where

$$v_k^c(r) = a_k^c r^k + b_k^c \frac{1}{r^k}, \quad v_k^s(r) = a_k^s r^k + b_k^s \frac{1}{r^k}.$$

Additionally, we denote the Fourier expansion of v_ρ on Γ_ρ by

$$v_\rho = c_0 + \sum_{k=1}^{\infty} \left(a_k \cos k\varphi + b_k \sin k\varphi\right) \tag{36}$$

From boundary conditions on $\Gamma_{\lambda\rho}$ it follows easily $B^v = 0, A^v = c_0$, and then $B^w = 0$, $A^w = A^v = c_0 = C_0$. There remains to find a_k, b_k, a_k^c, b_k^c, a_k^s, b_k^s, A_k^c, B_k^c, A_k^s, B_k^s assuming A_k, B_k as given.

5.3 Asymptotic Expansion

In order to eliminate the above-mentioned coefficients we consider first the terms at $\cos k\varphi$. From boundary and transmission conditions we have for $k = 1, 2, \ldots$

$$
\begin{aligned}
A_k^c R^k + B_k^c \frac{1}{R^k} &= A_k \\
A_k^c \rho^k + B_k^c \frac{1}{\rho^k} - a_k &= 0 \\
a_k^c \rho^k + b_k^c \frac{1}{\rho^k} - a_k &= 0 \\
a_k^c (\lambda\rho)^{k-1} - b_k^c \frac{1}{(\lambda\rho)^{k+1}} &= 0 \\
k_1 A_k^c \rho^{k-1} - k_1 B_k^c \frac{1}{\rho^{k+1}} - k_2 a_k^c \rho^{k-1} + k_2 b_k^c \frac{1}{\rho^{k+1}} &= 0
\end{aligned}
\tag{37}
$$

This may be rewritten in the matrix form. By grouping unknown parameters into a vector $\mathbf{p}_k = [A_k^c, B_k^c, a_k^c, b_k^c, a_k]^\top$ we obtain

$$
T(k_1, k_2, R, \lambda, \rho)\mathbf{p}_k = R^k A_k \mathbf{e}_1
$$

where

$$
T = \begin{bmatrix}
R^{2k} & 1 & 0 & 0 & 0 \\
\rho^{2k} & 1 & 0 & 0 & -\rho^k \\
0 & 0 & (\lambda\rho)^{2k} & 1 & -\rho^k \\
0 & 0 & (\lambda\rho)^{2k} & -1 & 0 \\
k_1\rho^{2k} & -k_1 & -k_2\rho^{2k} & k_2 & 0
\end{bmatrix}
\tag{38}
$$

where $\mathbf{e}_1 = [1,0,0,0,0]^\top$. It is easy to see that

$$
\mathbf{p}_k = \mathbf{p}_k^0 A_k + \rho^{2k}\mathbf{p}_k^1 A_k + o(\rho^{2k}) \tag{39}
$$

where

$$
\mathbf{p}_k^0 = \lim_{\rho\to 0+}\lim_{\lambda\to 0+} \frac{\mathbf{p}_k(k_1, k_2, R, \lambda, \rho)}{A_k}
$$

and $\mathbf{p}_k^0 = [1/R^k, 0, 0, 0, 0]^\top$, which corresponds to the ball B_R filled completely with material k_1. Similar reasoning may be conducted for terms containing $\sin k\varphi$.

As a result,

$$
\mathcal{A}_{\lambda,\rho} = \mathcal{A}_{0,0} + \rho^2 \mathcal{A}^1_{\lambda,\rho}(k_1, k_2, R, \lambda, \rho, A_1, B_1) + o(\rho^2). \tag{40}
$$

The exact form of $\mathcal{A}^1_{\lambda,\rho}(k_1, k_2, R, \lambda, \rho, A_1, B_1)$ is obtained from inversion of matrix T, but, what is crucial, it is linear in both A_1 and B_1. They, in turn, are computed as line integrals

$$A_1(w) = \frac{1}{\pi R^2}\int_{\Gamma_R} wx_1\,ds, \quad B_1(w) = \frac{1}{\pi R^2}\int_{\Gamma_R} wx_2\,ds.$$

As a result, for computing u_ρ we may use the following energy form

$$\begin{aligned}\mathcal{E}(u_\rho) = \frac{1}{2}\int_\Omega k_1 \nabla u_\rho \cdot \nabla u_\rho\,dx + \\ + \rho^2 Q(k_1,k_2,R,\lambda,\rho,A_1,B_1) + o(\rho^2),\end{aligned} \tag{41}$$

where $A_1 = A_1(u_\rho)$, $B_1 = B_1(u_\rho)$ and Q is a quadratic function of A_1, B_1. This constitutes a regular perturbation of the energy functional which allows computing perturbations of any functional depending on this solution and caused by small inclusion of the described above form.

5.4 Extension to Linear Elasticity

Let us consider the plane elasticity problem in the ring $C_{R,\rho}$. We use polar coordinates (r,θ) with $\mathbf{e}_r$ pointing outwards and $\mathbf{e}_\theta$ perpendicularly in the counterclockwise direction. Then there exists an exact representation of both solutions, using the complex variable series. It has the form [12, 17, 19]

$$\begin{aligned}\sigma_{rr} - i\sigma_{r\theta} &= 2\Re\phi' - e^{2i\theta}(\bar{z}\phi'' + \psi') \\ \sigma_{rr} + i\sigma_{\theta\theta} &= 4\Re\phi' \\ 2\mu(u_r + iu_\theta) &= e^{-i\theta}(\kappa\phi - z\bar{\phi}' - \bar{\psi}).\end{aligned} \tag{42}$$

The functions ϕ, ψ are given by complex series

$$\begin{aligned}\phi &= A\log(z) + \sum_{k=-\infty}^{k=+\infty} a_k z^k \\ \psi &= -\kappa\bar{A}\log(z) + \sum_{k=-\infty}^{k=+\infty} b_k z^k.\end{aligned} \tag{43}$$

Here μ—the Lame constant, ν—the Poisson ratio, $\kappa = 3-4\nu$ in the plain strain case, and $\kappa = (3-\nu)/(1+\nu)$ for plane stress.

Similarly as in the simple case described in former sections, the displacement data may be given in the form of Fourier series,

$$2\mu(u_r + iu_\theta) = \sum_{k=-\infty}^{k=+\infty} A_k e^{ik\theta} \tag{44}$$

The traction-free condition on some circle means $\sigma_{rr} = \sigma_{r\theta} = 0$. From (42), (43) we get the formula for displacements

$$2\mu(u_r + iu_\theta) = 2\kappa A r \log(r)\frac{1}{z} - \bar{A}\frac{1}{r}z + \\ + \sum_{p=-\infty}^{p=+\infty}[\kappa r a_{p+1} - (1-p)\bar{a}_{1-p}r^{-2p+1} \\ - \bar{b}_{-(p+1)}r^{-2p-1}]z^p. \tag{45}$$

Similarly, we obtain representation of tractions on some circle

$$\sigma_{rr} - i\sigma_{r\theta} = 2A\frac{1}{z} + (\kappa+1)\frac{1}{r^2}\bar{A}z + \\ + \sum_{p=-\infty}^{p=+\infty}(1-p)[(1+p)a_{p+1} + \bar{a}_{1-p}r^{-2p} \\ + \frac{1}{r^2}b_{p-1}]z^p. \tag{46}$$

As we see, in principle it is possible to repeat the same procedure again, glueing solutions in two rings together and eliminating the intermediary Dirichlet data on the interface. The only difference lies in considerably more complicated calculations, see, e.g., [9]. This could be applied for making double asymptotic expansion, in term of both ρ and λ. However, in our case λ does not need to be small in comparison with ρ.

6 Asymptotic Expansions of the Steklov–Poincaré Operators and Perturbations of Bilinear Forms in Particular Cases

The explicit form of solutions in B_R allows us to conclude that for

$$\|w_\rho\|_{H^{1/2}(\Gamma_R)} \leq \Lambda_0$$

the correction to the energy functional contains the part proportional to ρ^d and the remainder of order $o(\rho^d)$. This in turn [27, 28] implies the possibility of representation

$$w_\rho = w_0 + \rho^2 q + o(\rho^2) \quad \text{in} \quad H^1(\Omega_R)$$

for both standard and contact problems, justifying computations of topological derivatives.

It is well known that the singularities of solutions to Partial Differential Equations due to the singularities of geometrical domains can be characterized by specific shape derivatives of the associated energy shape functionals [7]. Therefore, the influence of topological changes in domains on the singularities can be measured by the appropriate second-order topological derivatives of the energy functionals. It means that we evaluate the shape derivatives of the energy functional by using the velocity field method, and subsequently the second-order topological derivatives of the functionals by an application of the domain decomposition method

- the portion Γ_0 of the boundary with the homogeneous Dirichlet boundary conditions is deformed to obtain $t \to T_t(V)(\Gamma_0)$ as well as $t \to \mathcal{E}(\Omega_t)$ for the energy shape functional; as a result the first order shape derivative $J(\Omega) := d\mathcal{E}(\Omega; V)$ is obtained in the distributed form as a volume integral,
- the second-order derivative of the energy functional is evaluated with respect to small parameter $\varepsilon \to 0$, the parameter governs the size of small inclusion with the material defined by a contrast parameter $\gamma \in [0, \infty)$.

We consider the energy shape functional $\Omega \to \mathcal{E}(\Omega)$ for Signorini problems for the Laplacian as well for the frictionless contact. The shape derivative $J(\Omega) := d\mathcal{E}(\Omega; V)$ of this functional is evaluated with respect to the boundary variations of the portion $\Gamma_0 \subset \partial\Omega$. In another words the velocity vector field V is supported in a small neighbourhood of Γ_0. The topological derivatives of $J(\Omega)$ are evaluated with respect to nucleation of small inclusions far from Γ_0. The domain decomposition method is applied in order to obtain the robust expressions for topological derivatives.

7 Directional Differentiability of the Metric Projection onto Positive Cone in Fractional Sobolev Spaces

Let us consider the subdomain $\Omega_c := \Omega \setminus \overline{\Omega}_R$ with the contact zone Γ_c in the scalar case as well as in an elastic body, see Fig. 1.

We recall that the convex cone for the contact problem in elasticity with linearized non-penetration conditions takes the form

$$\mathbb{K} := \{v \in H^1(\Omega_c) \ : \ [\![v]\!] \ \in \mathcal{K}(\Gamma_c) \subset H^{1/2}(\Gamma_c)\},$$

where $\mathcal{K}(\Gamma_c)$ is the positive cone in the fractional Sobolev space $H^{1/2}(\Gamma_c)$. The particular case is the space $H^{1/2}_{00}(\Gamma_c)$ for $\Gamma_c \subset \Sigma$ and the homogeneous Dirichlet conditions on the complement $\Sigma \setminus \overline{\Gamma}_c$, for the cracks. Therefore, we establish the *Hadamard differentiability* [11, 18] of the *metric projection* in the *Dirichlet space* $H^{1/2}(\Gamma_c)$ onto its *positive cone* [7].

Let us consider the directional differentiability of the metric projection onto the positive cone in the fractional Sobolev spaces $H^{1/2}(\Gamma_c)$. In order to present the

results, we are going to consider a simple geometry of the contact zone Γ_c. In the general setting the results can be obtained in the similar way. Therefore, we consider the subset $B = \{|x| < R\}, x = (x_1, \cdots, x_d) \subset \Omega$, of an elastic body Ω, with the contact set $\Gamma_c := \{x = (x', x_d) \in \mathbb{R}^d \ : \ x_d = 0, \quad |x'| < R/2\}$ and Σ defined by an extension of the subset $\widetilde{\Sigma} := \{x = (x', x_d) \in B \ : \ x_d = 0\}$. In such a case, the unit normal vector to the contact set $n := (0, \ldots, 0, 1)$ is constant on Γ_c, and the unit tangent vector orthogonal to n on the boundary $\partial\Gamma_c$ is $n := (n_1, \ldots, n_{d-1}, 0)$. For the displacement field $u = (u_1, \cdots, u_d)$ it follows that $un = u_d$, hence, the unilateral constraints for the normal component over the contact set $H^{1/2}_{00}(\Gamma_c) \ni [\![u]\!] n = u_d \geq 0$. Thus, the convex cone of admissible displacements for the contact problem takes the form

$$\mathcal{U}_{\rm ad} = \{v = (v_1, \cdots, v_d) \in H^1(\Omega_c) \ : \ v_d \geq 0 \ \text{ on } \ \Gamma_c\}$$

and our analysis of the metric projection is reduced to the positive cone in $H^{1/2}_{00}(\Gamma_c)$, hence, in $H^{1/2}(\Sigma)$.

Remark 3. We recall that in general for a domain Ω with the boundary Γ, the Sobolev spaces $H^1(\Omega)$ and $H^{1/2}(\Gamma)$ are [2, 10] the so-called Dirichlet spaces. It means that for the scalar product $a(\cdot, \cdot)$, with $v^+ := \sup\{v, 0\}$ and $v^- := \sup\{-v, 0\}$, the property $a(v^+, v^-) \leq 0$ holds for all elements of the Sobolev spaces.

Remark 4. The metric projection in the Dirichlet space onto the cone of nonnegative elements is considered for the purpose of sensitivity analysis of solutions to frictionless contact problems in [29]. This result is extended to the crack problem. In order to avoid unnecessary technicalities, we restrict ourselves to a model problem. Now, we consider the Hadamard differentiability of metric projection in Dirichlet space onto the cone of positive elements, and recall the result on its conical differentiability.

Consider the convex, closed cone

$$K = \{v \in H^{1/2}(\Sigma) \ : \ v \geq 0 \ \text{ on } \ \Sigma\}$$

and the metric projection $H^{1/2}(\Sigma) \ni f \to u = P_K(f) \in K$ onto K which is defined by the variational inequality

$$u \in K \ : \ (u - f, v - u)_{1/2,\Sigma} \geq 0 \quad \forall v \in K.$$

We denote $v^+ = v \wedge 0 := \sup\{v, 0\}$ and $v^- = -v \wedge 0 := \sup\{-v, 0\}$ in $H^{1/2}(\Sigma)$.

With the element $u = P_K(f)$ we associate the convex cone

$$C_K(u) = \{v \in H^{1/2}(\Sigma) \ : \ u + tv \in K \ \text{ for some } \ t > 0\}$$

and denote by $T_K(u)$ the closure of $C_K(u)$ in $H^{1/2}(\Sigma)$. On the other hand, [7] there is a nonnegative Radon measure m such that for all $v \in H^{1/2}(\Sigma)$ we have the equality $\int v\, dm = (u - f, v)_{1/2,\Sigma}$, hence, we denote

$$m[v] := (u-f,v)_{1/2,\Sigma}.$$

Definition 1. The convex cone K is polyhedric [11, 18] at $u \in K$ if

$$T_K(u) \cap m^{\perp} = \overline{C_K(u) \cap m^{\perp}}.$$

We recall the result on polyhedricity of the positive cone in a Dirichlet space [7].

Lemma 1. *The convex cone*

$$C_K(u) \cap m^{\perp} := \{v \in H^{1/2}(\Sigma) \ : \ v \in C_K(u) \ \ such \ \ that \ \ (u-f,v)_{1/2,\Sigma} = 0\}$$

is dense in the closed, convex cone

$$T_K(u) \cap m^{\perp} := \{v \in H^{1/2}(\Sigma) \ : \ v \in T_K(u) \ \ such \ \ that \ \ (u-f,v)_{1/2,\Sigma} = 0\}.$$

Proof. Using the property of the Dirichlet space

$$(v^+,v^-)_{1/2,\Sigma} \leq 0 \ \text{ for all } \ v \in H^{1/2}(\Sigma)$$

then

$$T_K(u) \cap m^{\perp} = \overline{C_K(u) \cap m^{\perp}}$$

follows easily.

Indeed, let

$$w \in T_K(u) \cap m^{\perp} \ .$$

Then $w = 0$ m-a.e. Let $C_K(u) \ni v_n \to w$. Then $v_n^- \to w^-$, $v_n^+ \to w^+$ and $v_n^+ \wedge w^+ - v_n^- \to w$, here $v \wedge w = \inf\{v,w\}$. Now, if $v \in C_K(u)$, then $u + tv \geq 0$. We claim $v_n^+ \wedge w^+ - v_n^- \in C_K(u) \cap m^{\perp}$. Indeed, $u + t[v_n^+ \wedge w^+ - v_n^-] \geq 0$ so $v_n^+ \wedge w^+ - v_n^- \in C_K(u)$ and $m[v_n^+ \wedge w^+ - v_n^-] = m[v_n^+ \wedge w^+] = 0$, because of $m[w^+] = 0$.

Remark 5. In [7] the tangent cone $T_K(u)$ is derived for $u \in K$, in the case of the positive cone $K = \{v \in \mathcal{H} \ : \ v \geq 0\}$ in the Dirichlet space $\mathcal{H}$ equipped with the scalar product $(u,v)_{\mathcal{H}}$. We have

$$T_K(u) = \{v \in \mathcal{H} \ : \ v \geq 0 \ \text{ on } \ \{u = 0\}\}.$$

The convex cone $S := T_K(u) \cap m^{\perp}$ is important for our applications. It is obtained in [7]

$$T_K(u) \cap m^{\perp} = \{v \in \mathcal{H} \ : \ v \geq 0 \ \text{ on } \ \{u = 0\} \ \text{ and } \ v = 0 \, m - \text{a.e.}\}.$$

The following result on the directional differentiability of metric projection holds for polyhedric convex sets [11, 18].

Lemma 2. *Let K be a polyhedric cone. For $t > 0$, t small enough,*

$$P_K(u+th) = P_K(u) + tP_S(h) + o(t;h) \ \ in \ \ H^{1/2}(\Sigma)$$

where

$$S := T_K(u) \cap m^{\perp}$$

and the remainder $o(t;h)$ is uniform on compact subsets of $H^{1/2}(\Sigma)$. Hence, the directional derivative of the metric projection is uniquely determined by the variational inequality

$$q := P_S(h) \in S \ : \ (q-h, v-q)_{1/2,\Sigma} \geq 0 \quad \forall v \in S.$$

For a contact set $\Gamma_c \subset \Sigma$ we introduce the following convex cones

$$\mathcal{K}(\Sigma) := \{v \in H^{1/2}(\Sigma) \ : \ v = 0 \ \text{ on } \ \Sigma \setminus \overline{\Gamma}_c, \quad v \geq 0 \ \text{ on } \ \Gamma_c\},$$

and

$$\mathcal{K}(\Gamma_c) := \{v \in H_{00}^{1/2}(\Gamma_c) \ : \ v \geq 0 \ \text{ on } \ \Gamma_c\}.$$

For the variational problems with unilateral conditions for the normal component of the displacement vector field over the contact set, the convex cones $\mathcal{K}(\Gamma_c)$ and $\mathcal{K}(\Sigma)$ are employed in order to show the polyhedricity of the cone of admissible displacements.

Remark 6. The proof of Lemma 1 applies as well to the convex cone $\mathcal{K}(\Gamma_c) \subset H_{00}^{1/2}(\Gamma_c)$ since the space $C_0^{\infty}(\Gamma_c)$ is dense in $H_{00}^{1/2}(\Gamma_c)$, hence, a nonnegative distribution is a Radon measure. In addition, contraction operates [4] for the scalar product (55) in $H_{00}^{1/2}(\Gamma_c)$. Let us note that the scalar products in $H^{1/2}(\Sigma)$ and in $H_{00}^{1/2}(\Gamma_c)$ are not the same, the latter is a weighted space.

We recall an abstract result on shape sensitivity analysis of variational inequalities.

Sensitivity Analysis of Variational Inequalities The conical differentiability of solutions to variational inequalities for the contact problem follows from the abstract result given by Theorem 1. The general result [29] is adapted here to our setting within the domain decomposition framework. Thus, the bilinear form $a(\cdot,\cdot) + b_t(\cdot,\cdot)$ defined in the subdomain Ω_c is introduced, where $b_t(\cdot,\cdot)$ is the contribution from the

Steklov–Poincaré operator on $\Gamma_R = \partial\Omega_R$. The real parameter $t > 0$ governs the shape perturbations of the inclusion $t \to \omega_t$ in Ω_R, where $t \to 0$ governs the topological changes of Ω_R in the framework of asymptotic analysis.

Two boundary value problems in two subdomains are coupled by the transmission conditions on the interface Γ_R. The linear boundary value problem in Ω_R furnishes the expansions of the Steklov–Poincaré operators resulting from perturbations of the inclusion in the interior of the subdomain. The sensitivity analysis of solutions to variational inequality in Ω_c is performed for compact perturbations of nonlocal boundary conditions on the interface. As a result, the weak solution to the unilateral elasticity boundary value problem under considerations is directionally differentiable with respect to the parameter $t \to 0$ which governs the perturbations of the inclusion far from the contact set.

Now, we provide the precise result on the conical differentiability of solutions to variational inequalities [11, 18, 29] (see also [7]) which is given here without the proof.

Let $\mathcal{K} \subset \mathcal{H}$ be a convex and closed subset of a Hilbert space $\mathcal{H}$, and let $\langle \cdot, \cdot \rangle$ denote the duality pairing between $\mathcal{H}'$ and $\mathcal{H}$, where $\mathcal{H}'$ denotes the dual of $\mathcal{H}$. Let us assume that there are given symmetric bilinear forms $a(\cdot,\cdot) + b_t(\cdot,\cdot) : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ parametrized by $t \geq 0$, and the linear form $f \in \mathcal{H}'$, such that

Condition 1. 1. *There are* $0 < \alpha \leq M$ *such that*

$$|a(u,v) + b_t(u,v)| \leq M\|u\|\|v\|, \quad \alpha\|u\|^2 \leq a(v,v) + b_t(v,v) \quad \forall u, v \in \mathcal{H} \tag{47}$$

uniformly with respect to $t \in [0, t_0)$. *Furthermore, there exists* $\mathcal{Q}' \in \mathcal{L}(\mathcal{H}; \mathcal{H}')$ *such that*

$$\mathcal{Q}_t = \mathcal{Q} + t\mathcal{Q}' + o(t) \quad \textit{in} \quad \mathcal{L}(\mathcal{H}; \mathcal{H}')\,, \tag{48}$$

where $\mathcal{Q}_t \in \mathcal{L}(\mathcal{H}; \mathcal{H}')$

$$a(\phi, \varphi) + b_t(\phi, \varphi) = \langle \mathcal{Q}_t(\phi), \varphi \rangle \quad \forall \phi, \varphi \in \mathcal{H}.$$

2. *The set* $\mathcal{K} \subset \mathcal{H}$ *is convex and closed, and the solution operator* $\mathcal{H}' \ni f \to \mathcal{P}(f) \in \mathcal{H}$ *for* (52)

$$\mathcal{P}(f) \in \mathcal{K}: \quad a(\mathcal{P}(f), \varphi - \mathcal{P}(f)) \geq \langle f, \varphi - \mathcal{P}(f) \rangle \quad \forall \varphi \in \mathcal{K} \tag{49}$$

is differentiable in the sense that

$$\forall h \in \mathcal{H}': \quad \mathcal{P}(f + sh) = \mathcal{P}(f) + s\mathcal{P}'(h) + o(s) \quad \textit{in} \quad \mathcal{H} \tag{50}$$

for $s > 0$, s small enough, where the mapping $\mathcal{P}' : \mathcal{H}' \to \mathcal{H}$ is continuous and positively homogeneous, in addition, the remainder $o(s)$ is uniform with respect to the direction $h \in \mathcal{H}'$ on compact subsets of $\mathcal{H}'$.

Let us consider the unique solutions $u_t = \mathcal{P}_t(f)$ to variational inequalities depending on a parameter $t \in [0, t_0)$, $t_0 > 0$,

$$u_t \in \mathcal{K} \ : \ a(u_t, \varphi - u_t) + b_t(u_t, \varphi - u_t) \geq \langle f, \varphi - u_t \rangle \qquad \forall \varphi \in \mathcal{K} \, . \tag{51}$$

In particular, for $t = 0$

$$u \in \mathcal{K} \ : \ a(u, \varphi - u) + b(u, \varphi - u) \geq \langle f, \varphi - u \rangle \qquad \forall \varphi \in \mathcal{K} \, , \tag{52}$$

with $u = \mathcal{P}(f)$ a unique solution to (52). The mapping $t \to u_t$ is strongly differentiable in the sense of Hadamard at 0^+, and its derivative is given by a unique solution of the auxiliary variational inequality [29].

Theorem 1. *Assume that Condition 1 is satisfied. Then the solutions to the variational inequality (51) are right-differentiable with respect to t at $t = 0$, i.e. for $t > 0$, t small enough,*

$$u_t = u + tu' + o(t) \quad \text{in} \quad \mathcal{H} \, , \tag{53}$$

where

$$u' = \mathcal{P}'(-\mathcal{Q}'u) \, . \tag{54}$$

7.1 Metric Projection onto Positive Cone in $H_{00}^{1/2}(\Gamma_c)$

For boundary value problems in domains with contact conditions, unilateral conditions are prescribed on the contact set for the normal component of the displacement field. Hence, the normal component of the displacement field belongs to the positive cone in the fractional Sobolev space $H^{1/2}(\Gamma_c)$. The sensitivity analysis of variational inequalities for Signorini problems was reduced in [29] to the directional differentiability of the metric projection onto the positive cone in a fractional space which is the Dirichlet space. This result is further extended in [7] to some crack problem. The method is also used in the present paper, however for the purposes of sensitivity analysis of contact problems.

Sensitivity Analysis of the Crack Problem We are going to explain how the results obtained in [29] for the Signorini problem in linear elasticity can be extended to the crack problems with unilateral constraints. To this end, the abstract analysis performed in [7] for the differentiability of the metric projection onto the cone of nonnegative elements in the Dirichlet space is employed.

The framework for analysis is established in function spaces over $\Omega := \Omega^+ \cup \Sigma \cup \Omega^-$, where Σ is a $C^{1,1}$ regular curve without intersections. The regularity assumption can be weakened, if necessary.

Let $\Gamma_c \subset \Sigma$ be the segment $\{(x_1,0) \ : \ 0 < x_1 < 1\}$ included in the curve Σ. We denote by n the unit normal vector field on Σ which points out of Ω^+, and by τ the unit normal vector field on $\partial\Gamma_c$ orthogonal to n. We consider deformations of the crack in the direction of the vector field V collinear with τ in the neighbourhood of the crack tip $A = (1,0) \in \Omega_c \subset \mathbb{R}^2$.

In the Sobolev space defined on the cracked domain Ω_c, the elements enjoy jumps over the crack which are denoted by $[\![v]\!] := v^+ - v^-$, and we have the regularity property of traces $[\![v]\!] \in H_{00}^{1/2}(\Gamma_c)$. In our geometry of Ω_c, the Sobolev space $H_{00}^{1/2}(\Gamma_c)$ coincides with the linear subspace of $H^{1/2}(\Sigma)$

$$H_{00}^{1/2}(\Gamma_c) = \{\varphi \in H^{1/2}(\Sigma) \ : \ \varphi = 0 \quad \text{q.e. on } \Sigma \setminus \Gamma_c\},$$

where q.e. means *quasi-everywhere* with respect to the capacity, see, e.g., [1, 24] for the definition and elementary properties of the capacity useful for the existence of optimal shapes in shape optimization problems with nonlinear PDE's constraints.

In order to investigate the properties of the metric projection in the space of admissible displacement fields onto the convex cone

$$K := \{v \in H^1(\Omega_c) \ : \ [\![v]\!]\, n \geq 0\},$$

where $H^1(\Omega_c) := H^1(\Omega_c;\mathbb{R}^2)$, we need to show that the positive convex cone

$$\mathcal{K} = \{\varphi \in H_{00}^{1/2}(\Gamma_c) \ : \ \varphi \geq 0 \quad \text{on } \Gamma_c\}.$$

is polyhedric in the sense of [7, 11, 18].

We consider here the rectilinear crack Γ_c in two spatial dimensions. The scalar product in $H_{00}^{1/2}(\Gamma_c) := H_{00}^{1/2}(0,1)$ is defined

$$\langle \varphi, \psi \rangle_c = \int_{\Gamma_c} \int_{\Gamma_c} \frac{(\varphi(x) - \varphi(y))(\psi(x) - \psi(y))}{|x-y|^2} dxdy \tag{55}$$
$$+ \int_{\Gamma_c} \left[\varphi(x)\psi(x) + \frac{\varphi(x)\psi(x)}{\mathrm{dist}(x, \partial\Gamma_c)} \right] dx$$

Polyhedricity of the Positive Cone in $H_{00}^{1/2}(\Gamma_c)$ In order to show the polyhedricity of the nonnegative cone $\mathcal{K}$ in $\mathcal{H} := H_{00}^{1/2}(0,1)$, it is enough to check the property

$$\langle \varphi^+, \varphi^- \rangle_c \leq 0 \quad \forall v \in H_{00}^{1/2}(0,1)$$

which is straightforward, here $\varphi^+(x) = \max\{v(x), 0\}$. The full proof of polyhedricity in such a case is provided in [7]. It is easy to check that the polyhedricity with respect to the scalar product implies the polyhedricity with respect to a bilinear form which is equivalent to the scalar product.

Theorem 2. *Let us consider the variational inequality for the metric projection of* $f + th \in \mathcal{H}$ *onto* $\mathcal{K}$

$$u_t \in \mathcal{K} \ : \ \langle u_t - f - th, v - u_t \rangle \geq 0 \quad \forall v \in \mathcal{K},$$

where $f, h \in \mathcal{H}$ *are given, denote by* $\Xi\{u\} = \{x \in \Gamma_c \ : \ u(x) = 0\}$. *Then*

$$u_t = u + tq(h) + o(t;h) \quad in \quad \mathcal{H},$$

where the remainder $o(t;h)$ *is uniform on compact subsets of* $\mathcal{H}$, *and the conical differential of the metric projection* $q := q(h)$ *is given by the unique solution to the variational inequality*

$$q \in \mathcal{S}(u) \ : \ \langle q - h, v - q \rangle \geq 0 \quad \forall v \in \mathcal{S}(u)$$

and the closed convex cone

$$\mathcal{S}(u) = \{\varphi \in \mathcal{H} \ : \ \varphi \geq 0 \quad q.e. \ on \quad \Xi\{u\}, \quad \langle u - f, \varphi \rangle = 0\}.$$

8 Rectilinear Crack in Two Spatial Dimensions

In this section the general method of shape-topological sensitivity analysis is presented in the domain $\Omega := \Omega_c \cup \Gamma_R \cup \Omega_R$, where the first subdomain Ω_c contains the rectilinear cracks Γ_c and the second subdomain Ω_R contains the inclusion ω.

We denote by $\Omega_{\text{in}} := \Omega_c \cup \overline{\Gamma}_c$, the first subdomain in the elastic body without the crack. We assume that there is a regular $C^{1,1}$-curve $\Sigma \subset \Omega_{\text{in}}$, without intersections, which contains the rectilinear crack $\Gamma_c := \{(x_1, 0) \ : \ 0 \leq x_1 \leq 1\}$. To simplify the presentation, let us consider a torus $\Omega := \mathbb{T} := \mathbb{T}^2$ with 2π-periodic coordinates $x = (x_1, x_2)$.

The deformations of the subdomain Ω_c are defined by the vector field $(x,t) \to V(x,t) = (v(x,t), 0)$, where the $C_0^\infty(\Omega^+)$ function $(x,t) \to v(x,t)$ is supported in $[1-\delta, 1+\delta]^2 \times [-t_0, t_0] \subset \Omega^+ \subset \mathbb{R}^2 \times \mathbb{R}$ and $v(x,t) \equiv 1$ on $[1-\delta/2, 1+\delta/2]^2 \times [-t_0/2, t_0/2]$. In our notation, the real variable $t \in \mathbb{R}$ is a parameter. It means that the vector field V deforms the reference domain Ω_c^+ to $t \to T_t(V)(\Omega_c^+)$ just by moving the tip of the crack $X = (1,0) \to x(t) = (x_1(t), 0)$ in the direction of the x_1-axis. The mapping $T_t \ : \ X \to x(t)$ is given by the system of equations

$$\frac{dx}{dt}(t) = V(x(t),t), \quad x(0) = X.$$

The boundary value problem of linear isotropic elasticity in Ω_c is defined by the variational inequality

$$u \in K \ : \ a(u, v-u) \geq (f, v-u) \quad \forall u \in K, \tag{56}$$

where

$$K = \{v \in H^1(\Omega_c) \ : \ [\![v]\!] \cdot n := (v^+ - v^-) \cdot n \geq 0 \ \text{ on } \ \Gamma_c\}, \tag{57}$$

here $[\![v]\!] = v^+ - v^-$ is the jump of the displacement field over the crack Γ_c. The bilinear form

$$a(u,v) = \int\limits_{\Omega_c} \left[\frac{\mu}{2} \sum_{j,k=1}^{2} (\partial_j u_k + \partial_k u_j)(\partial_j v_k + \partial_k v_j) + \lambda \operatorname{div} u \operatorname{div} v \right] dx$$

is associated with the operator

$$Lu := -\mu \Delta u - (\lambda + \mu) \mathbf{grad} \operatorname{div} u. \tag{58}$$

The deformation tensor $2\varepsilon(u) = \partial_j u_k + \partial_k u_j$ and the stress tensor $\sigma(u)$ associated with the displacement field u are useful in the description of the boundary value problems in linear elasticity.

The energy functional $\mathcal{E}(\Omega_c) = 1/2 a(u,u) - (f,u)_{\Omega_c}$ is twice differentiable [7] in the direction of a vector field V, for the specific choice of the field $V = (v, 0)$. The first order shape derivative

$$V \to d\mathcal{E}(\Omega_c; V) = \frac{1}{t} \lim_{t \to 0} (\mathcal{E}(T_t(\Omega_c)) - \mathcal{E}(\Omega_c)$$

can be interpreted as the derivative of the elastic energy with respect to the crack length, we refer the reader to [15] for the proof, the same result for the Laplacian is given in [5, 13, 14].

Theorem 3. *We have*

$$d\mathcal{E}(\Omega_c; V) = \frac{1}{2} \int\limits_{\Omega_c} \left\{ \operatorname{div} V \cdot \varepsilon_{ij}(u) - 2E_{ij}(V;u) \right\} \sigma_{ij}(u) - \int\limits_{\Omega_c} \operatorname{div}(V f_i) u_i \, . \tag{59}$$

Now we restrict our consideration to the perturbation of the crack tip only in the direction which coincides with the crack direction. The derivative is evaluated in the framework of the velocity method [29] for a specific velocity vector field V selected in such a way that the result $d\mathcal{E}(\Omega_c; V)$ is independent of the field V and

it depends only on the perturbation of the crack tip. That is why, this derivative is called the *Griffith's functional* $J(\Omega_c) := d\mathcal{E}(\Omega_c; V)$ defined for the elastic energy in a domain with crack. We are interested in the dependence of this functional on domain perturbations far from the crack. As a result, shape and topological derivatives of the nonsmooth Griffith's shape functional are obtained with respect to the boundary variations of an inclusion.

8.1 Green Formulae and Steklov–Poincaré Operators

The Steklov–Poincaré operator on the interface for the domain $\Omega_c \cup \Gamma_R \cup \Omega_R$ is defined by the Green formula, first as the Dirichlet-to-Neumann map in Ω_R, then it is used on the interface as nonlocal boundary operator. Therefore, we recall here the Green formula for linear elasticity operators in two and three spatial dimensions.

We start with analysis in two spatial dimensions. To simplify the presentation let us consider the reference domain without a crack in the form of the torus $\mathbb{T} := \mathbb{T}^2$ with 2π-periodic coordinates $x = (x_1, x_2)$. For the purpose of shape-topological sensitivity analysis we assume that the elastic body without the crack is decomposed into two subdomains, Ω_{in} and Ω_R, separated from each other by the interface Γ_R. Thus, the elastic body with the crack Γ_c is written as

$$\Omega := \Omega_c \cup \Gamma_R \cup \Omega_R \,.$$

The rectilinear crack $\Gamma_c \subset \Sigma \subset \Omega_{in}$ is an open set, where the fictitious interface $\Sigma \subset \Omega_{in}$ is a closed $C^{1,1}$-curve without intersections. In our notation $\Omega_c = \Omega_{\text{in}} \setminus \overline{\Gamma}_c$.

The bilinear form of the linear isotropic elasticity is associated with the operator

$$Lu := -\mu \Delta u - (\lambda + \mu) \mathbf{grad} \operatorname{div} u$$

for given Lame coefficients $\mu > 0, \lambda \geq 0$.

The displacement field u in the elastic body Ω is given by the unique solution of the variational inequality

$$u \in K \ : \ a(u, v-u) \geq (f, v-u) \quad \forall u \in K \,, \tag{60}$$

where

$$K = \{v \in H^1(\Omega_c) \ : \ [\![v]\!] \cdot n := (v^+ - v^-) \cdot n \geq 0 \ \text{ on } \ \Gamma_c\}, \tag{61}$$

here $[\![v]\!] = v^+ - v^-$ is the jump of the displacement field over the crack Γ_c.

Given the unique solution $u \in K$ of the variational inequality and the admissible vector field V compactly supported in Ω_c, we consider the associated shape functional (59) evaluated in Ω_c, which is called the Griffith's functional

$$J(\Omega_c) := d\mathcal{E}(\Omega_c; V) \,. \tag{62}$$

Let $\omega \subset \Omega_R$ be an elastic inclusion. Introduce the family of inclusions $t \to \omega_t \subset \Omega_R$ governed by the velocity field W compactly supported in Ω_R. The elastic energy in Ω_R with the inclusion ω_t is denoted by

$$\omega_t \to \mathcal{E}_t(\Omega_R) := \frac{1}{2} a_t(\Omega_R \ ; \ u,u) - (f,u)_{\Omega_R}.$$

Its shape derivative $d\mathcal{E}(\Omega_R; W)$ in the direction W is obtained by differentiation of the function at $t = 0$

$$t \to \mathcal{E}_t(\Omega_R) := \frac{1}{2} a_t(\Omega_R \ ; \ u,u) - (f,u)_{\Omega_R}.$$

Proposition 4. *Assume that the energy shape functional in the subdomain Ω_R,*

$$\omega \to \mathcal{E}(\Omega_R) := \frac{1}{2} a(\Omega_R \ ; \ u,u) - (f,u)_{\Omega_R}$$

is differentiable in the direction of the velocity field W compactly supported in a neighbourhood of the inclusion $\overline{\omega} \subset \Omega_R$, then the Griffith's functional (62) is directionally differentiable in the direction of the velocity field W. Therefore, the second-order directional shape derivative $d\mathcal{E}(\Omega; V, W)$ of the energy functional in Ω in the direction of fields V, W is obtained.

This result can be proved by the domain decomposition technique:

- the shape differentiability of the energy functional in the subdomain Ω_R implies the differentiability of the associated Steklov–Poincaré operator defined on the Lipschitz curve given by the interface $\overline{\Omega}_R \cap \overline{\Omega}_c$ with respect to the scalar parameter $t \to 0$ which governs the boundary variations of the inclusion ω;
- the expansion of the Steklov–Poincaré nonlocal boundary pseudodifferential operator obtained in the subdomain Ω_R is used in the boundary conditions for the variational inequality defined in the cracked subdomain Ω_c and leads to the conical differential of the solution to the unilateral problem in the subdomain;
- the one term expansion of the solution to the unilateral problem is used in the Griffith's functional in order to obtain the directional derivative with respect to the boundary variations of the inclusion.

Remark 7. For the circular inclusion $\omega := \{x \in \Omega_R \ : \ |x-y| < r_0\}$, $r_0 > 0$, the scalar parameter $t \to 0$ which governs the shape perturbations of $\partial\omega$ in the direction of a field W [29] can be replaced by the parameter $r \to r_0$. Thus, the moving domain $t \to \omega_t$ is replaced by the moving domain $r \to \{x \in \Omega_R \ : \ |x-y| < r\}$. In this way the shape sensitivity analysis [29] for $r_0 > 0$ and the topological sensitivity analysis [22] for $r_0 = 0^+$ are performed in the same framework for the simple case of circular inclusion.

9 Shape and Topological Derivatives of Elastic Energy in Two Spatial Dimensions for an Inclusion

In the subdomain Ω_c the unique weak solutions

$$\varepsilon \to u := u_\varepsilon$$

of the elasticity boundary value subproblem are given by the variational inequality

$$u \in K \ : \ a(\Omega_c; u, v-u) + b_\varepsilon(\Gamma_R; u, v-u) \geq (f, v-u)_{\Omega_c} \quad \forall v \in K.$$

In order to differentiate the solution mapping for this variational inequality, it is required to differentiate the bilinear form $\varepsilon \to b_\varepsilon(\Gamma_R; u, v)$, which is performed in this section.

9.1 Shape and Topological Derivatives of the Energy Functional in Ω_R with Respect to the Inclusion ω

In order to evaluate the topological derivative of energy functional in isotropic elasticity, the shape sensitivity analysis is combined with the asymptotic analysis [22], see also [8, 16, 23] for related results. In this section the small parameter is denoted by $\varepsilon \to 0$, and the circular inclusion $\varepsilon \to \omega_\varepsilon := B_\varepsilon$ is considered.

The general shape of inclusion $\varepsilon \to \omega_\varepsilon$ can be considered in the same way for shape sensitivity analysis [29] and the asymptotic analysis [22].

For the sake of simplicity, the subscript R is omitted, thus, we denote $\Omega := \Omega_R$, since the inclusion is located in the subdomain Ω_R. We also allow for the Neumann Γ_N and Dirichlet Γ_D pieces of the boundary $\partial\Omega := \partial\Omega_R$, thus, $\partial\Omega_R := \Gamma_N \cup \Gamma_D \cup \Gamma$. Thus, we evaluate the shape and topological derivative [22] of the total potential energy associated with the plane stress linear elasticity problem, considering the nucleation of a small inclusion, represented by $B_\varepsilon \subset \Omega$, as the topological perturbation. In this way the expansion of the Steklov–Poincaré operator on the interface $\Gamma := \Gamma_R$ is obtained.

Steklov–Poincaré Operator Let us consider the nonhomogeneous Dirichlet linear elasticity boundary value problem in the domain Ω with the boundary $\partial\Omega := \Gamma_N \cup \Gamma_D \cup \Gamma$.

$$\left\{ \begin{array}{l} \text{Find } u, \text{ such that} \\ \operatorname{div}\sigma(u) = 0 \qquad \text{in } \Omega\,, \\ \sigma(u) = \mathbb{C}\nabla u^s\,, \\ u = 0 \qquad \text{on } \Gamma_D\,, \\ u = \overline{u} \qquad \text{on } \Gamma\,, \\ \sigma(u)n = 0 \qquad \text{on } \Gamma_N\,, \end{array} \right. \tag{63}$$

where the only nonhomogeneous term is the Dirichlet condition $u = \overline{u}$ on the interface Γ. Let

$$a(u,u) := \int_\Omega \sigma(u) \cdot \nabla u^s$$

stands for the associated bilinear form, thus the elastic energy of the solution u is given by

$$\mathcal{E}(\Omega;u) = \frac{1}{2} a(u,u).$$

Then by Green's formula

$$\mathcal{E}(\Omega;u) = \langle \mathcal{T}(\overline{u}), \overline{u} \rangle_\Gamma. \tag{64}$$

In the case of an inclusion $\omega_\varepsilon \subset \Omega$, the formula becomes

$$\mathcal{E}_\varepsilon(\Omega;u) = \langle \mathcal{T}_\varepsilon(\overline{u}), \overline{u} \rangle_\Gamma. \tag{65}$$

Hence, the expansion of the energy functional in Ω, on the left-hand side of (65) with respect to the parameter $\varepsilon \to 0$ can be used in order to determine the associated expansion of the Steklov–Poincaré operator $\overline{u} \to \mathcal{T}(\overline{u})$ on the right-hand side of (65).

Acknowledgements This work has been supported by the DFG EC315 "Engineering of Advanced Materials" and by the ANR-12-BS0-0007 Optiform.

References

1. Adams, D., Helberg, L.: Function Spaces and Potential Theory. Springer, Berlin (1996)
2. Ancona, A.: Sur les espaces de Dirichlet: principes, fonction de Green. J. Math. Pures Appl. **54**, 75–124 (1975)
3. Argatov, I.I., Sokolovski, Y.: Asymptotics of the energy functional in the Signorini problem under small singular perturbation of the domain. (Russian. Russian summary) Zh. Vychisl. Mat. Mat. Fiz. **43**(5), 744–758 (2003); translation in Comput. Math. Math. Phys. **43**(5), 710–724 (2003)
4. Beurling, A., Deny, J.: Dirichlet spaces. Proc. Natl. Acad. Sci. U.S.A. **45**, 208–215 (1959)
5. Delfour, M.C., Zolésio, J.-P.: Shapes and Geometries. SIAM, Philadelphia, PA (2001)
6. Evans, L.C.: Partial Differential Equations. Graduate Studies in Mathematics, vol. 19, 2nd edn. American Mathematical Society, Providence, RI (2010)
7. Frémiot, G., Horn, W., Laurain, A., Rao, M., Sokołowski, J.: On the analysis of boundary value problems in nonsmooth domains. Diss. Math. **462**, 149 (2009)
8. Garreau, S., Guillaume, P., Masmoudi, M.: The topological asymptotic for PDE systems: the elasticity case. SIAM J. Control Optim. **39**, 1756–1778 (2001)
9. Gross, W.A.: The second fundamental problem of elasticity applied to a plane circular ring. Z. Angew. Math. Phys. **8**, 71–73 (1957)

10. Hanouzet, B., Joly, J.-L.: Méthodes d'ordre dans l'interprétation de certaines inéquations variationnelles et applications. J. Funct. Anal. **34**, 217–249 (1979)
11. Haraux, A.: How to differentiate the projection on a convex set in Hilbert space. Some applications to variational inequalities. J. Math. Soc. Jpn. **29**, 615–631 (1977)
12. Kachanov, M., Shafiro, B., Tsukrov, I.: Handbook of Elasticity Solutions. Kluwer Academic Publishers, Dordrecht (2003)
13. Khludnev, A.M., Sokołowski, J.: Griffith's formula and Rice-Cherepanov's integral for elliptic equations with unilateral conditions in nonsmooth domains. In: Optimal Control of Partial Differential Equations (Chemnitz, 1998). International Series of Numerical Mathematics, vol. 133, pp. 211–219. Birkhäuser, Basel (1999)
14. Khludnev, A.M., Sokołowski, J.: The Griffith's formula and the Rice-Cherepanov integral for crack problems with unilateral conditions in nonsmooth domains. Eur. J. Appl. Math. **10**, 379–394 (1999)
15. Khludnev, A.M., Sokołowski, J.: Griffith's formulae for elasticity systems with unilateral conditions in domains with cracks. Eur. J. Mech. A. Solids **19**, 105–119 (2000)
16. Lewinski, T., Sokołowski, J.: Energy change due to the appearance of cavities in elastic solids. Int. J. Solids Struct. **40**, 1765–1803 (2003)
17. Lurie, A.I.: Theory of Elasticity. Springer, Berlin/Heidelberg (2005)
18. Mignot, F.: Contrôle dans les inéquations variationelles elliptiques. J. Funct. Anal. **22**, 130–185 (1976)
19. Muskhelishvili, N.I.: Some Basic Problems on the Mathematical Theory of Elasticity. Noordhoff, Groningen (1952)
20. Nazarov, S.A., Sokołowski, J.: Asymptotic analysis of shape functionals. J. Math. Pures Appl. **82**(2), 125–196 (2003)
21. Nazarov, S.A., Sokołowski, J.: Self-adjoint extensions for the Neumann Laplacian and applications. Acta Math. Sin. (Engl. Ser.) **22**(3), 879–906 (2006)
22. Novotny, A.A., Sokołowski, J.: Topological Derivative in Shape Optimization. Springer, Berlin (2013)
23. Novotny, A.A., Feijóo, R.A., Padra, C., Taroco, E.A.: Topological sensitivity analysis. Comput. Methods Appl. Mech. Eng. **192**, 803–829 (2003)
24. Plotnikov, P., Sokołowski, J.: Compressible Navier-Stokes Equations: Theory and Shape Optimization. Mathematics Institute of the Polish Academy of Sciences. Mathematical Monographs (New Series), vol. 73. Birkhäuser/Springer Basel AG, Basel (2012)
25. Sokołowski, J., Żochowski, A.: On the topological derivative in shape optimization. SIAM J. Control Optim. **37**(4), 1251–1272 (1999)
26. Sokołowski, J., Żochowski, A.: Optimality conditions for simultaneous topology and shape optimization. SIAM J. Control Optim. **42**(4), 1198–1221 (2003)
27. Sokołowski, J., Żochowski, A.: Modelling of topological derivatives for contact problems. Numer. Math. **102**(1), 145–179 (2005)
28. Sokołowski, J., Żochowski, A.: Topological derivatives for optimization of plane elasticity contact problems. Eng. Anal. Bound. Elem. **32**(11), 900–908 (2008)
29. Sokołowski, J., Zolésio, J.-P.: Introduction to Shape Optimization, Shape Sensitivity Analysis. Springer Series in Computational Mathematics, vol. 16. Springer, Berlin (1992)

Optimal Control for Applications in Medical and Rehabilitation Technology: Challenges and Solutions

Katja Mombaur

Abstract This paper gives an overview of the mathematical background and possible applications of optimal control and inverse optimal control in the field of medical and rehabilitation technology, in particular in human movement analysis, therapy and improvement by means of appropriate medical devices. One particular challenge in this area is the formulation of suitable subject-specific models of motions for healthy and impaired humans including skeletal multibody dynamics and potentially neuromuscular components, and their combination with models of the technical components. The formulation of hybrid multi-phase optimal control problems arising in this context involves non-standard elements such as the open- or closed loop stability of the dynamic motion. Efficient methods for the solution of optimal control and inverse optimal control are discussed and particular difficulties of this problem class are highlighted. In addition, we present several example applications of these methods in the development of mobility aids for geriatric patients, the optimization-based design of exoskeletons, the analysis of running motions with prostheses, the optimal functional electrical stimulation of hemiplegic patients, as well as stability studies for different types of movement.

1 Introduction

Optimal control problems are ubiquitous in medical applications, in particular in the field of motion generation and analysis in rehabilitation. This paper serves to highlight challenges for applied mathematics, and especially optimization and optimal control, arising from this interesting area of applications and to present potential solution approaches. The goals in this context are to better understand human movement and to use this understanding to improve the movement either by controlling it directly, by developing better training and rehabilitation techniques, or by optimally designing technical devices that support or guide the movement.

K. Mombaur (✉)
Interdisciplinary Center for Scientific Computing (IWR), Optimization in Robotics and Biomechanics, Heidelberg University, Berliner Str. 45, 69120 Heidelberg, Germany
e-mail: katja.mombaur@iwr.uni-heidelberg.de

J.-B. Hiriart-Urruty et al. (eds.), *Advances in Mathematical Modeling, Optimization and Optimal Control*, Springer Optimization and Its Applications 109,
DOI 10.1007/978-3-319-30785-5_5

The human is a complex, dynamic, unpredictable system with a control system of its own, and the particular challenge is to make the complex technology developed nevertheless work in conjunction with the human in a variety of situations.

It is a common assumption that many processes and structures in nature are optimal. This optimality hypothesis also holds for natural motions of humans and animals that have been shaped by evolution, learning, and training [3, 4]. From a mathematical perspective it is therefore logical to formulate dynamic motion tasks such as walking, running, standing up, grasping, etc. as optimal control problems. Optimality therefore appears in two different ways in this context—first, as a way to predict natural human movement, and second to increase performance of the technical devices.

The first goal of this paper is to discuss in detail some of the mathematical challenges arising from these applications and their potential solutions, such as:

- The efficient and flexible modeling of these complex biomechanical systems: The mathematical descriptions of such motions result in highly nonlinear systems of ordinary differential or differential-algebraic equations, generally including multiple phases of motion, implicitly defined phase changes and discontinuities of state variables between phases. These multibody system models need to be adjustable to different subjects and situations, and the right level of complexity has to be chosen for each application. The identification of good data for human models also presents a big issue.
- Neuromuscular modeling: The problem gets even significantly more complex if muscular elements and neural excitation and control loops are considered in the model. This has so far only been done for parts of the human body, and the establishment of a whole-body human model with muscles and neural parts for excitation and control that can efficiently be used in forward simulation and optimization is subject to ongoing and future work.
- Integrating models of the technical devices: since the interaction of the human with the medical device is to be investigated, detailed models of the devices and control systems are required and must be integrated with the human model in a combined model.
- A correct formulation of optimal control problems for the generation and control of such motions: this generally results in a hybrid multi-phase optimal control problem including switches, continuous and discrete phases, constraints and objective functions. Avoiding global and local redundancy of the constraints poses a particular challenge in some applications. Objective functions can get very complex as soon as stability issues are involved: in this case, derivatives of the trajectories have to be considered in the objective function or constraint formulations, and the variational differential equation of the hybrid dynamics has to be included in the dynamic constraints of the problem.
- The initialization of state and control variables: due to the large number of variables an automated initialization is favorable, and due to the local nature of the optimization procedures, a good initialization is very important.

- An efficient solution of optimal control problems is essential, both for offline (generation/selection of motions) and online problems (control of motions). Direct optimal control techniques using multiple shooting have proved very efficient for the solution of such problems. While in the offline case, precise solutions for whole body human models can be sought for, reduced models and real-time methods have to be used in the online case. Also methods for model reduction play an important role in this context.
- An efficient solution of inverse optimal control problems: Inverse optimal control problems are formulated to identify optimization objectives of motions from (partial) measurements of state variables and potentially control variables. This class of problems is particularly challenging, since it consists in solving a parameter estimation problem in an optimal control problem. Bi-level as well as one-level methods have been developed to solve this type of problems.
- Handling of uncertainties and variability in data: data in this context is recorded by optical motion tracking systems, inertial measurement units, force plates, EMG, etc. None of these measurements is precise. In addition, there is a lot of variation between subjects, motion trials, scenarios, etc. Deciding which data can be combined for which analysis (e.g., which motions are combined in one inverse optimal control computation with the hypothesis that they share the same underlying objective function) is a very hard problem.
- The transfer of optimization results to reality also is an issue. Once optimal motions have been computed for a prosthesis, an exoskeleton, another physical assistive device, or a stimulation pattern, they have to be applied to the real system, and methods for coping with the model mismatch are required.

The second goal of this paper is to summarize some of the work on medical optimal control problems performed in our research group on different topics in each of which several of the above challenges have been successfully addressed. We will present optimal control problems in the following applications:

- optimization of geriatric motions with and without the support of physical assistive device, in particular optimal sit to stand transfer;
- the study of walking and running motions with prostheses, e.g. in the context of disability sports: high-speed running of bipedal amputees on carbon fiber prostheses;
- the optimization-based development and control of an exoskeleton for full motion support of paraplegic patients;
- the generation of optimal muscle stimulation patterns for functional electrical stimulation (FES) in walking, in particular for the treatment of the drop foot syndrome of stroke patients;
- the analysis and improvement of stability of normal and pathological gait.

The paper is organized as follows: Sect. 2 describes the types of mathematical models that are required for motion studies. In Sect. 3, we give an overview of different mathematical stability criteria. Section 4 presents our approach to the formulation and solution of optimal control problems to generate motions for this

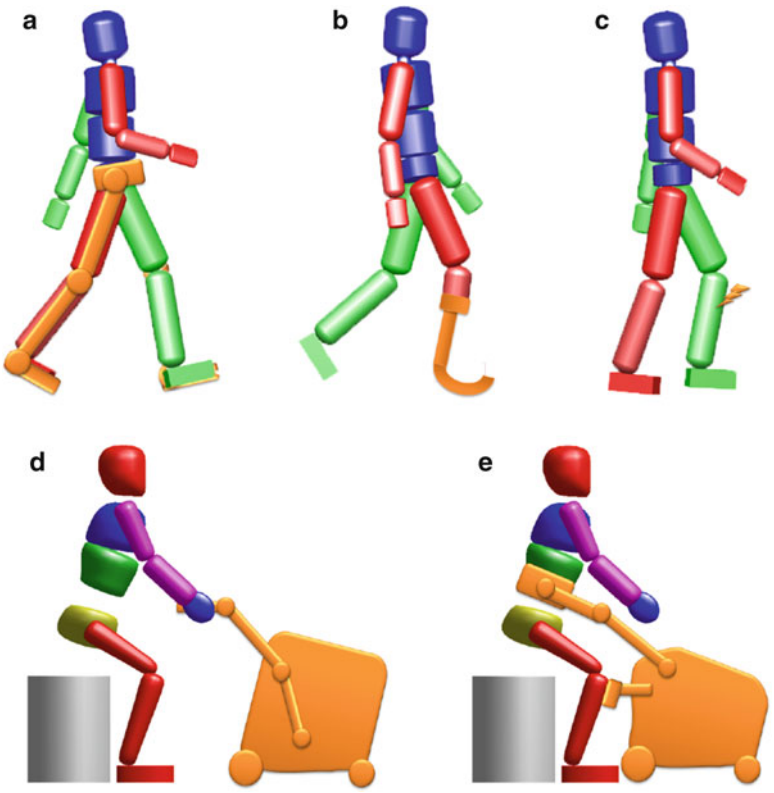

Fig. 1 Example applications from medical technology and rehabilitation discussed in this paper in which model-based optimal control has been applied: (**a**) Lower-limb exoskeleton (**b**) Running prosthesis (**c**) Functional electrical stimulation of the tibialis anterior (**d**) Rollator type actuated physical assistive device (**e**) Nurse type physical assistive device

type of applications. In Sect. 5 we introduce inverse optimal control problems. Sections 6–10 summarize results of optimal control projects in the five medical application areas listed in the previous paragraph. In Sect. 11, we draw some conclusions and give an outlook to future optimization work in that area.

2 Mathematical Models for Motion Studies in Medical and Rehabilitation Engineering

This section discusses challenges related to the formulation of mathematical models of humans and technical devices for optimization-based motion studies and medical and rehabilitation applications. From the mathematical perspective, it should not be expected that human whole-body models in a form that is numerically suitable

for optimization are available from the medical or biomechanical field. Existing models are quite often too simple or sometimes mechanically wrong, and they typically include discontinuities and non-differentiabilities. In the past, we have therefore developed in our group two different modeling tools which represent general multibody systems modeling tools but serve in our case the primary goal to generate whole-body human and humanoid models. The first one, named RBDL, [19, 20], is based on an order n recursive algorithm by Featherstone [18], the other one, named Dynamod [38], is based on explicit code generation. The choice of the level of detail of the model is not straightforward and depends on the particular question asked. We will discuss different choices in the context of the applications in Sects. 6–10. All models involve many anthropomorphic parameters. Model data is highly individual and very hard to identify. Some data can be measured (but measurements may be expensive), whereas others can only be estimated, and there is no way to measure them. We will also comment on some data choices within this section and the sections on the different applications.

In this section, we first present rigid body models that describe the whole-body actions of the human assuming that there are torques acting on the joints. After that, we discuss what would have to be added to the model in order to describe how these torques result from the actions of the neuromuscular system. In the end of this section, we describe how models of medical and rehabilitation devices can be set up and integrated with the human models. Due to lack of space, none of these sections is meant to give a complete overview of the respective area. Instead, we try to give a flavor of the challenges and of the types of models that are to be faced, and present some solution approaches.

2.1 Whole-Body Models of Humans

To describe the human body we use a multibody system model with a set of rigid bodies that are connected by different types of joints. The choice of using three-dimensional (3D) models or planar (2D) models (e.g., in the sagittal plane) as well as the precise number of degrees of freedom (DOF) depends on the particular question investigated. For our investigations we use 3D models with 35–40 DOF in some cases, and 2D models with 9–15 DOF in others. In the general case, we assume that the system is powered by torques in the internal joints. For geometry and inertia parameters, tabular anthropometric data can be used, e.g. from de Leva [42] or Winter [75], but if a precise match between model and real human or good prediction quality is to be achieved, the parameters have to be adjusted to the individual properties of the subject or at least of a particular group (see, e.g., Sect. 6). One approach to establish subject-specific dynamic human models based on kinematic measurements is proposed in [20, 21].

The motions that we consider in our research generally consist of multiple phases of motion produced by the changing contacts of the human with the environment, in many cases by his hands and feet, and discontinuities between phases, i.e. they take

the form of hybrid dynamical problems. Each phase is characterized by its own set of equations of motion. Depending on the choice of coordinates q, we obtain ordinary differential or differential algebraic equations for each phase, all of them highly nonlinear. For the choice of minimal coordinates q (i.e., the number of coordinates is equal to the number of DOF of the system), the motion is described by a set of ordinary differential equations of the following form:

$$M(q,p)\ddot{q} + N(q,\dot{q},p)\dot{q} = F(q,\dot{q},p,\mathscr{M}), \tag{1}$$

M is the mass or inertia matrix and N the vector of nonlinear effects. F is the vector of all external forces (including gravity, joint torques $\mathscr{M}$, drag, etc.) Note that $q(t)$ and $v(t) = \dot{q}(t)$ are functions in time and form the state variables of the system, and p is the vector of model parameters which are fixed in time, but may still be free parameters in the optimal control problems to be discussed in the next sections.

If redundant coordinates q are used (i.e., there are more position variables than DOF), the coupling can be described by a constraint of the form $g(q,p) = 0$ and a corresponding constraint force in the differential equation. This results in a system of differential algebraic equations (DAE) of index 3 for the equations of motion. For numerical treatment, we reduce this to a DAE of index 1 by index reduction:

$$\dot{q} = v \tag{2}$$

$$\dot{v} = a \tag{3}$$

$$\begin{pmatrix} M(q,p) & G(q,p)^T \\ G(q,p) & 0 \end{pmatrix} \begin{pmatrix} a \\ \lambda \end{pmatrix} = \begin{pmatrix} -N(q,v) + F(q,v,p,\mathscr{M}) \\ \gamma(q,v,p) \end{pmatrix} \tag{4}$$

$$g_{pos} = g(q(t),p) = 0 \tag{5}$$

$$g_{vel} = G(q(t),p) \cdot \dot{q}(t) = 0. \tag{6}$$

with acceleration $a = \ddot{q}$ and Lagrange multipliers λ. The matrix G is the Jacobian of the position constraints $G = (\partial g/\partial q)$, and γ the corresponding Hessian $\gamma = ((\partial G/\partial q)\,\dot{q})\,\dot{q}$. Equations (5) and (6) describe the invariants on position and velocity level resulting from the index reduction that the solution still must satisfy.

In the case of modeling tools that are based on explicit code generation, such as Dynamod [38] or HuMAnS [74], we obtain codes for the different parts of these equations, such as M, G, N, and γ which allows us to set up the above systems of equations and solve them for the accelerations a. In the case of modeling tools based on recursive formalisms, such as RBDL [19], we don't obtain an explicit code for these expressions, but only the resulting accelerations based on a recursive evaluations for the kinematic tree structure of the system. No unique answer can be given to the question which of the two approaches is preferable in terms of computational speed since it depends on the precise characteristics of the system, e.g. the number of DOF and the particular processor used as well as the specific implementation. However it can be stated that computation times for both codes developed in our group (RBDL and Dynamod) are in the same order of magnitude for the type of problems discussed here and are internationally very competitive.

Several contacts occurring in motion models, such as ground contact during walking and running, are unilateral (i.e. the ground can only push against the foot but not pull it). This can be taken care of in optimization by formulating an appropriate inequality constraint on the Lagrange multiplier associated with the normal contact force (see Eq. (4)).

Phase changes between the motion phases described above usually do not occur at given time points, but depend on the states of the system, which can be described by a phase switching condition of type

$$s(q(\tau_s), v(\tau_s), p) = 0. \tag{7}$$

To give an example from walking and running motions: touch-down occurs when the lowest point of the foot reaches ground height, and lift-off takes place when the vertical contact force (i.e., the negative of the corresponding Lagrange multiplier in Eq. (4)) becomes zero.

The discontinuities between phases that make up the hybrid nature of the problem are usually discontinuities of the velocities that are caused by inelastic impacts (e.g., at touchdown) which instantaneously set velocities at some points to zero. The other velocities after impact v_+ can be computed as

$$\begin{pmatrix} M(q,p) & G(q,p)^T \\ G(q,p) & 0 \end{pmatrix} \begin{pmatrix} v_+ \\ \Lambda \end{pmatrix} = \begin{pmatrix} M(q)v_- \\ 0 \end{pmatrix}, \tag{8}$$

where v_- are the corresponding velocities immediately before impact. Matrices M and G are the same as in Eq. (4).

Some of the motions that we consider, like walking and running motions, are periodic or quasi-periodic. It is therefore often suitable to impose periodicity constraints to the model on all velocity variables v and a reduced set of position variables q_{red}, only eliminating the coordinate describing the person's direction of motion. Typically, in this case, we are also interested to determine a symmetric gait with identical left and right steps, such that the periodicity constraint is applied after one step, after formulating a shift of sides in the model.

2.2 Including Neuromuscular Models in the Human Model

In the models described in the previous section, motions are driven by joint torques $\mathscr{M}$. For many medical applications it may be desirable or even necessary to explore the process of joint torque generation further and include models of the underlying mechanisms into the overall model of human movement. These models can include the process of force generation in the muscles as well as all processes related to neural excitation of muscles and to the feedback of sensory signals about motion, muscle states, and the environment.

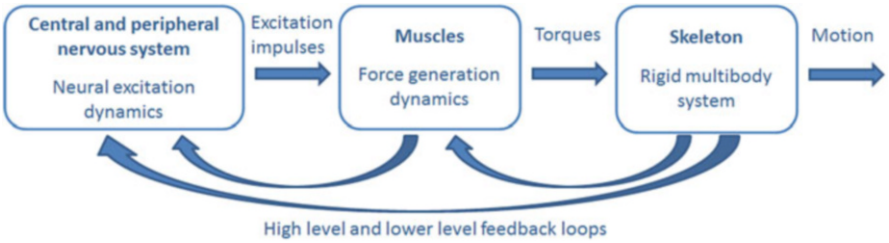

Fig. 2 Components of a full neuro-musculo-skeletal human model

An overview of the different components of a human neuro-musculo-skeletal model is given in Fig. 2. While the box to the right was treated in the previous section, here we discuss the boxes in the middle and briefly the one to the left. It is important to note at this point that the state of the art is still very far from the formulation of a full dynamic human neuro-musculo-skeletal model. Some attempts for setting up musculo-skeletal models (without the neural part) have been pursued successfully (e.g., [14, 59, 60]). We will now first discuss some details about muscle modeling in the next paragraphs, and then come back to the issue of neuronal modeling towards the end of this section.

In humans and animals, joint torques M_i at joint i result from the sum of torques M_{ij} produced by all $n_{mus,i}$ muscles acting on the respective joint which are the product of the muscle forces F_{ij} times the lever arms d_{ij}:

$$M_i = \sum_{j=1}^{n_{mus,i}} M_{ij} = \sum_{j=1}^{n_{mus,i}} F_{ij} \cdot d_{ij}. \tag{9}$$

Muscle models are used to describe forces of individual muscles F_{ij} as a function of the input and of the system's state. There are various types of muscle models. The most popular class of models is formed by the so-called Hill type muscle models that go back to the original work of Hill [30]. There are many variants of the Hill type model in use today. The model we describe here corresponds essentially to the model established by van den Bogert and Ackermann [1, 11]; one exception is that we assume here a constant tendon length while they allow for variable tendon length which is important for some muscles. We give here the equations to give an idea about the mathematical form of these models.

As in most Hill-type models, the muscle is modeled as a combination of a contractile element, a parallel elastic and a parallel damping element $F_{TA} = F_{CE} + F_{PE} + F_{PD}$. The force in the contractile element is generally computed as the product of the maximum isometric force $F_{iso,max}$ and three factors, the activation level f_{ad}, the force-length factor f_{fl}, and the force-velocity factor f_{fv}:

$$F_{CE} = F_{iso,max} \cdot f_{ad} \cdot f_{fl} \cdot f_{fv} \tag{10}$$

The activation level f_{ad} (with $0 \leq f_{ad} \leq 1$) which can be seen as an input of the above equation is at the same time the output of the activation dynamics of the muscle

$$\dot{f}_{ad} = (\epsilon - f_{ad}) \left(\frac{\epsilon}{T_{act}} - \frac{1.0 - \epsilon}{T_{deact}} \right). \tag{11}$$

Muscle excitation ϵ ($0 \leq \epsilon \leq 1$) is the entry or control variable of this model of activation dynamics, and T_{act} and T_{deact} are the time constants for activation and deactivation of the particular muscle.

The length of the muscle-tendon complex can be computed as a function of the corresponding joint angles. The computation depends obviously on the fact if the muscle spans one or two joints. In the case of monoarticular muscles it is of the form

$$l_{MT} = l_{MT,0} \pm d_l \cdot \phi \tag{12}$$

with rest length $l_{MT,0}$, muscle specific constant d_l, relative joint angle ϕ, and the sign of the last term depending on the sign convention for the joint angle ϕ and the side of the joint on which the muscle acts. The tendon length l'_T is assumed to be constant here and the length of the contractile element, i.e. the muscle itself, follows from

$$l_{CE} = l_{MT} - l_T. \tag{13}$$

In this case, the contraction speed of the muscle can be computed as

$$v_{CE} = \pm d_l \cdot \dot{\phi}. \tag{14}$$

If the tendon is modeled as elastic, typically a linear spring model is used, and tendon length and muscle length have to be iteratively determined from the force balance between muscle and tendon.

The force-length factor f_{fl} describes the dependency of the muscle force on the current length of the contractile element l_{CE}

$$f_{fl} = e^{-\left(\frac{l_{CE} - l_{CE,opt}}{W l_{CE,opt}}\right)^2} \tag{15}$$

where $l_{CE,opt}$ denotes the fiber length at which the optimum force can be generated, and W is the so-called width parameter that describes how the filaments in the sarkomer overlap which is the crucial low level mechanism responsible for the force length relationship.

The force-velocity factor f_{fv} describes the dependency of the muscle force on the contraction speed of the muscle and can be modeled by two hyperbolic relationships:

$$f_{fv} = \frac{g_{max} v_{CE} + c}{v_{CE} + c} \quad \text{if } v_{CE} > 0 \text{ (extension)}, \tag{16}$$

$$f_{fv} = \frac{\lambda v_{max} + c}{\lambda v_{max} - v_{CE}/A} \quad \text{else (contraction)}, \tag{17}$$

where g_{max} is the normalized maximum force during extension and A is a constant. With the parameter λ, we take into account that the activation level influences the maximum contraction speed

$$\lambda(f_{ad}) = 1 - e^{-c_1 f_{ad}} + f_{ad} e^{-c_1}. \tag{18}$$

The factor c in Eqs. (16) and (17) is introduced to produce continuous first order derivatives at $v_{CE} = 0$ where the two branches intersect, and is computed as

$$c = \frac{\lambda v_{max} A (g_{max} - 1)}{A + 1}. \tag{19}$$

The force in the parallel elastic element can be computed as

$$F_{PE} = k_1 (l_{CE} - l_{slack,PE}) \quad \text{if } l_{CE} \leq l_{slack,PE}, \tag{20}$$

$$F_{PE} = k_1 (l_{CE} - l_{slack,PE}) + k_{2,PE} (l_{CE} - l_{slack,PE})^2 \quad \text{else}, \tag{21}$$

while the force in the parallel damping element is

$$F_{DE} = b v_{CE}. \tag{22}$$

These equations contain several parameters, such as d_l, T_{act}, T_{deact}, W, A, k_1, $l_{slack,PE}$, b,c, c_1, which are different for each muscle and for each person. Determining good parameter values for muscle models is a very difficult issue, since some parameters can only be identified in cadavers and therefore no truly individual values can be obtained for a living subject.

We would like to point out that the whole set of Eqs. (10)–(22) is necessary to describe the dynamics of one single muscle. As stated in (9), several muscles—at least two muscles, an agonist and an antagonist, but often more—are responsible for the torque around one joint axis. So to give an idea of how much complexity a decision to include muscle models into the whole-body model brings along: this means that for a human model with 30 internal DOF, we would need to add at least 60 muscles, i.e. add 60 times the equations stated in this paragraph to the previous system described in Sect. 2.1. And for all these muscles, appropriate model parameters would have to be determined and good input variables chosen. However, for several medical applications this additional effort must be taken, since it is important to study muscle behavior in detail to explore constraints, delays, or interfaces with technical devices.

Even more challenging is the formulation of good models of the neuronal control of muscle activity including all feedback loops. Muscle excitation processes for movement generation are far from being fully understood with some signals coming from the brain, and others from the spinal cord. No full model exists yet, but attempts to set up neuro-muscular models for part of the body such as the arm have been established, e.g. in [70]. It is assumed that the neural control does not work with the full set of mechanical degrees of freedom, but performs control tasks on a much smaller space [8]. In the previous paragraph it was still assumed that the incoming excitations of the different muscles of the whole-body model are independent. This is however not true, since higher level control levels of the CNS and the peripheral nervous system typically send commands to whole groups of muscles which are responsible for the same tasks. This is investigated under the name of muscle synergies or muscle excitation primitives by various researchers, e.g. [23, 69, 71].

The research field of Human motor control (e.g., [67]) is dedicated to the general study of how humans use their neuromuscular system to move their body and coordinate the limbs and how sensorimotor signal about the state of the body and the environment are integrated. However, a lot of research so far is performed on simpler motor skills such as pointing and reaching and no global models of the human body that would allow to investigate medical devices in the context of whole-body motions and address, e.g. stability questions, have been established yet.

2.3 Modeling the Rehabilitation Device and Establishing a Combined Model

A major challenge in the context of medical and rehabilitation applications is to design and control technical devices that optimally support humans with different pathologies and for different types of motions. While the particular demands for the different systems discussed in that paper—prostheses, orthoses and exoskeletons, external physical assistive devices and stimulation equipment- are different, the common request is that they are able to work in conjunction with the human patient and significantly improve his or her motion.

For this, it is essential to have an integrated model of the human and the respective device that allows to study their interactions and their movement as a whole and that allows to evaluate different design and control selections for the device. So in addition to the model discussed in the previous sections, the following steps have to be taken:

(a) Development of good mechanical models of the device;
(b) Choice of appropriate strategy of combination of device model with human model.

Part (a) obviously depends a lot on the particular device considered. What can be said in general:

- In most of the cases that we considered so far, devices are also modeled as multibody system models and resulted in further sets of ordinary differential equations or differential-algebraic equations of the same type as above (1)–(8).
- All device models are set up in a parameterized way in order to allow for parameter optimization in the optimal control context.
- Some devices are purely passive, i.e. they consist of springs, dampers, passive locking devices, etc. One example for this is the running prostheses discussed in Sect. 7 which essentially is a carbon fiber spring. Also soft tissues with different passive properties are gaining more importance in this field.
- Some devices are active and can include all kinds of linear or rotational actuators (electric, hydraulic, pneumatic, etc.). Depending on the particular question, there is the possibility to
  - either represent the actuators by their resulting forces or torques which then act as input to the remaining model,
  - or to include full models of the actuators themselves in order to take their dynamics into account which then results in additional differential equations. For example in the case of electric motors a model of the following type can be used:

$$\psi(i) = \tau(\dot{\varphi}) \cdot \left(a \cdot i - d_M(i) \right. - \frac{1}{2\pi} \kappa k_f \dot{\varphi} \left. - I \kappa \ddot{\varphi}\right), \tag{23}$$

    with τ being the motor's efficiency, $a \cdot i - d_M(i)$ describing the linear characteristics of electric motors with slope a and damping term d_M, the anti proportionality factor k_f of the torque to the number of revolutions per second, transmission ratio κ, and the inertia I of the rotor. These characteristic parameters would have to be determined for every motor.

- In some cases, an inclusion of partial differential equations may be very useful for describing some details of the device, e.g. if large deformable parts with contacting surfaces are involved, e.g. soft soles of an exoskeleton. This goes however beyond the scope of this paper and will not be discussed in the examples here.

With respect to (b), the goal is to integrate the model of the human (Sects. 2.1 and potentially 2.2) and the model of the medical device discussed in (a) into a combined model that allows to optimize the actions of the whole system at once. Different approaches maybe applied depending on the type of device and the amount of contact between human and device as well as on the level of detail desired:

- Rigid coupling of the medical device to the human, assuming perfect nonsliding contact between the two and perfect alignment of the joint axes (where kinematic structures overlap, e.g. in exoskeletons). This results in a multi-body system which consists of all segments of the two subsystems and in the case of overlapping structures of segments that are combinations of attached human and device (exoskeleton) segments. As results we obtain combined joint efforts and combined loads on the whole system.

- Compliant coupling between human and the device at several identified contact points by the formulation of spring elements. This generally results in larger systems of equations since human and device do not share any degrees of freedom, but allows to compute separate torques and loads on human and device.
- Compliant coupling between human and the device on extended compliant surfaces caused by soft tissue and/or compliant elements in the devices such as cushions, etc. Such a contact model again has to be formulated by means of finite elements that have to be integrated in the whole-body human model.

3 Mathematical Stability Criteria for Human Movement

Stability of human movement with or without supporting devices is a crucial property in many applications in medical technology and rehabilitation. Human motions are typically not statically stable (i.e., the center of mass is not sitting well within the polygon of support and the motion is not negligible) such that dynamic or at least quasi-dynamic stability criteria have to be used. We give an overview of some stability criteria in this section, and come back to applications of stability criteria in Sect. 10.

There is no uniquely accepted way to define stability in the context of motions, and we will discuss different possible criteria in this section.

3.1 ZMP Related Criteria

In the field of humanoid robots, stability definitions based on the concept of the zero moment point (ZMP) play an important role. The ZMP [73] defines the point where the resulting torques of inertia and gravity forces of the robot about the horizontal axes lying in the ground become zero. The ZMP is equivalent to the center of pressure (CoP) for nonsliding motions on level ground, but while the CoP is computed based on ground reaction forces, the ZMP is generally defined using inertia and accelerations of the segments of the robots.

$$p_{ZMP,Q} = \frac{n \times M_Q^{gi}}{R_{gi} \cdot n} \tag{24}$$

where $p_{ZMP,Q}$ denotes the position of the ZMP with respect to a general point Q, n denotes the normal direction of the contact force, R_{gi} the sum of the gravity and inertia force at the center of mass, and M_Q^{gi} the moment at point Q caused by acceleration, gravity, and change in angular momentum of the segments [68].

In humanoid robotics, for simplicity and speed, the ZMP is often computed using the simplified so-called table-cart model of Kajita [36] in which essentially only the horizontal movement of the pelvis center is considered and all the mass

is assumed to be gathered in this point. The position of the table cart ZMP on the floor $(p_{ZMP,x}, p_{ZMP,y})$ in forward (x) and sideward (y) direction then follows from the center of mass position x, y, z_c (with the height z_c assumed to be constant) and accelerations:

$$p_{ZMP,x} = x - \frac{z_c}{g}\ddot{x}, \tag{25}$$

$$p_{ZMP,y} = y - \frac{z_c}{g}\ddot{y}. \tag{26}$$

To produce stable gaits, ZMP control algorithms aim at keeping the simplified ZMP within the polygon of support, usually with a large safety margin to the boundary. Some of the most famous control algorithms for humanoid robots fall into that category. This results in quite conservative and slow gaits which do not look very dynamic. CoP measurements for human movement show that the CoP/ZMP goes to the edge of the polygon of support quite frequently: even during normal walking the CoP travels from extreme heel to toes during each foot contact. During very dynamic balancing motions, the CoP goes to the boundary of the polygon of support at many different places and for extended periods of time, and with the ZMP criterion, even if precisely computed, it is not possible to predict if the system will fail in the next second or if it can be stabilized. For medical applications, a ZMP criterion only can be used in very limited cases, where the motions are not very dynamic, and the basin of support is large and therefore large safety margins can be applied without constraining the motion too much.

3.2 Lyapunov's First Method

From a mathematical perspective, stability in the sense of Lyapunov is a much better way to describe stability of moving systems. Here stability is not defined in terms of a momentary glimpse on the position of a point, but by looking at the behavior of the solution under the effect of small perturbations.

It is well known that a solution of a non-autonomous nonlinear system is asymptotically stable in the sense of Lyapunov if small perturbations of the initial values result in a perturbed solution that always stays in a finite neighborhood of the original solution (stability) and if the effect of the perturbation vanishes for $t \to \infty$, i.e. the perturbed solution converges to the unperturbed one (e.g., [13]). For autonomous systems, we consider orbital asymptotic stability which corresponds to the above definition with the exception that orbital shifts of the solution by the perturbation (i.e., shifts in time) may occur and remain, but are not considered.

In human motions many stability questions arise in the context of periodic motions such as gaits, so we have to consider the stability of periodic limit cycles. This is defined by Lyapunov's first method (compare e.g. [13, 34]): a T-periodic solution of a T-periodic non-autonomous system

$$\dot{x}(t) = f(t, x(t)) \text{ with } f(t, \cdot) = f(t+T, \cdot) \tag{27}$$

is asymptotically stable if all eigenvalues of the monodromy matrix are inside the unit circle

$$|\lambda_i(X(T))| < 1. \tag{28}$$

The monodromy matrix—also referred to as transfer matrix or sensitivity matrix or Jacobian of the Poincaré map is defined as

$$X(T) = \frac{dx(T)}{dx(0)}, \tag{29}$$

i.e., it describes the sensitivities or first order derivatives of the end values of the trajectories with respect to their initial values.

If not all entries of x are periodic, e.g. if there is one direction of travel in the motion, the non-periodic directions have to be eliminated by projection prior to using the eigenvalue stability criterion. Also in the case of autonomous systems, i.e. systems without any actuation or other kinds of input variables, where there is always an invariant eigenvalue of one describing that perturbations along the orbit are conserved, this eigenvalue has to be eliminated by projection.

Note that this stability criterion is also valid for hybrid multi-phase systems, if the order of phases is preserved and the state after the discontinuity is twice continuously differentiable with respect to initial values (although non-differentiable in time), as we have shown in [53]. For a motion consisting of several phases, the monodromy matrix over the whole interval $[0, T]$ is computed as the matrix product of the matrices at the individual phases:

$$X(0, t_m = T) = X(t_{m-1}, t_m) \cdot \ldots \cdot X(t_1, t_2) \cdot X(0, t_1). \tag{30}$$

If a discontinuity $J(t, x)$ occurs at some point t_s that is only implicitly defined by a switching function s, then the so-called update formula (to update the monodromy matrix) has to be applied:

$$X(0, t_m = T) = X(t_s, t_m) \cdot U(t_s) \cdot X(0, t_s) \tag{31}$$

and the update of the matrix at the discontinuity is computed as

$$U(t_s) = \left(\Delta f - J_t - J_x f(t_s^-)\right) \cdot \frac{1}{\dot{s}} (s_x)^T + I + J_x \tag{32}$$

with Δf being the discontinuity in the right-hand side f, J_t and J_s the partial derivatives of J, s_x the partial derivative of the switching function f with respect to x, and $\dot{s}$ its total derivative with respect to time.

We will present how we have applied this criterion for stability optimization in the context of optimal control problems, see Sect. 4. This criterion has been applied to walking motions of very simple systems by various authors (e.g., [12, 24, 35, 47]), who used it to analyze the stability of a given motion, but not in optimization.

Other than the spectral radius, possible choices for objective functions are induced matrix norms of the monodromy matrix, such as the 1- or ∞-norm or the singular value which all are upper bounds on the spectral radius according to the theorem of Hirsch (e.g., Theorem 6.9.1 in [72]) and therefore represent stricter measures of stability .

3.3 Lyapunov's Second Method

The more famous second method of Lyapunov takes a different approach than the previously described first method (see many textbooks on differential equations, e.g. [13]. It does not rely on any linearization or first order sensitivity information, but instead uses a so-called Lyapunov function $V(t,x)$ to determine if the solution $x \equiv 0$ of the nonlinear differential equation $\dot{x} = f(t,x(t))$, i. e. $0 = f(t,0)$ is stable.

The concept of the Lyapunov function is inspired by the fact that the potential energy of a physical system is minimal at a stable equilibrium and maximal at an unstable equilibrium. The Lyapunov function $V(t,x)$ represents a generalization of the potential energy function. The Lyapunov function is defined on the domain $D_v = \{(t,x) | t > t_1, |x| < A\}$ and must have the following properties:

- continuous first partial derivatives with respect to t and x_i : $V(t,x) \in \mathscr{C}^1(D)$,
- $V(t,0) = 0$ for $t > t_1$,
- positive definiteness: $V(t,x) > 0$ for $x \neq 0$,
- negative definiteness of derivative: $\dot{V}(t,x) \leq 0$.

The derivative $\dot{V}(t,x)$ which is the derivative of $V(t,x)$ along the solution $x(t)$ is defined as

$$\dot{V}(t,x(t)) = \sum_{i=1}^{n} \frac{\partial V}{\partial x_i} \dot{x}_i + \frac{\partial V}{\partial t}. \tag{33}$$

Lyapunov's second method states that if such a function exists, then the trivial solution $0 = f(t,0)$ is stable. In detail, we distinguish the following cases:

- $\dot{V}(x) \leq 0$ in D $\Rightarrow$ stability in the sense of Lyapunov,
- $\dot{V}(x) < 0$ in D $\Rightarrow$ asymptotic stability,
- $\dot{V}(x) \leq -\alpha V(x)$ and $V(x) \geq b|x|^{\beta}$ in D with $(\alpha, \beta, b > 0)$ $\Rightarrow$ exponential stability.

If this method is to be applied to a real world application, e.g. in medical and rehabilitation technology, a suitable Lyapunov function for this system must be constructed. This is difficult, since such functions have only been found for certain classes of systems, e.g. the total energy is a Lyapunov function for Hamiltonian systems. In particular the use of stability criteria in stability optimization, as we

envision it (see next section) requires the automatic construction and evaluation of such a function in every optimization step which is very hard. In recent years, an approach to construct Lyapunov functions based on sums of squares has been proposed [61] and was also extended to systems with contacts and discontinuities [64]. So far, this method has been applied successfully to develop controllers for given equilibrium solution [46], but not to optimize the solution, i.e. the motion itself, as we wish to do it. This is still ongoing work.

3.4 Capture Point Stability Criteria

Another type of stability criteria that has become very popular in robotics is based on the capture point. The capture point is defined as the point on the floor where the human or robot would have to place its foot—or to be more precise, the CoP (or the ZMP)—in order to come to a full stop [66]. As in the case of the ZMP, the capture point reflections are usually based on a linear inverted pendulum, i.e. an inverted pendulum for which the mass stays at the same height. The capture point is also known in biomechanics under the name of "extrapolated center of mass" (Xcom) [33]. In order to reach the stop, the pendulum mass has to come to rest exactly above the COP which results in the following equation for the capture point (here in x direction, y direction is equivalent):

$$p_{CP,x} = x + \frac{\dot{x}}{\omega} \text{ with } \omega = \sqrt{\frac{g}{z_c}}. \tag{34}$$

The system dynamics in x direction can be described with the following two equations for the capture point and the center of mass:

$$\dot{x} = -\omega(x - p_{CP,x}) \tag{35}$$

$$\dot{p}_{CP,x} = \omega(p_{CP,x} - p_{ZMP,x}). \tag{36}$$

While the dynamics of the center of mass x are stable, those of the capture point $p_{CP,x}$ are unstable. In robotics, several capture point related control approaches are based on the concept of only controlling the capture point since the center of mass follows automatically, e.g. [17]. There are several possible extensions (see [40]):

- instead of the capture point, also the concept of n-step capturability and capture regions (stopping in n steps instead of 1 step) is explored;
- the linear inverted pendulum can be augmented by additional components, e.g. a finite sized foot instead of a point foot, or an extended mass instead of a point mass as upper body.

The capture point is certainly an interesting concept for rehabilitation applications, but the exact type of model still remains to be determined.

3.5 Angular Momentum

Another criterion that is frequently mentioned in the context of stability of human motion is the total angular momentum about the center of mass

$$H = \sum_{i=1}^{n} (r_i \times m_i \dot{r}_i) + \sum_{i=1}^{n} (\Theta_i \omega_i) \tag{37}$$

and its change. m_i and Θ_i are the mass and inertia of segment i, r_i the distance of the segment center of mass from the total center of mass, $\dot{r}_i$ the corresponding velocity, and ω_i the angular velocity. During walking and running motions as well as in any other upright form of movement without rotations the average angular momentum must be zero since there is no persisting rotation and the overall orientation of the human body remains the same. This changes for motions with rotations such as somersaults for which a significant angular momentum is generated about the frontal axis and for spinning jumps, e.g. in figure skating, where angular momentum about the vertical axis is required. Physically, angular momentum can only change if a external torque is applied, i.e. the angular momentum about the center of mass is constant about any possible axis when the human is in the air. During walking and running it can change by the action of ground reaction forces. The absolute values of the total angular momentum about the center of mass during walking as well as the contributions of the different segments have been studied by Popovic et al. [62] and Herr and Popovic [28]. The values are quite small over the cycle but show characteristic peaks and zeros. The result has also been confirmed in our research on emotional locomotion [20].

4 Formulation and Solution of Optimal Control Problems for Motion Generation

The dynamic models presented in Sect. 2 can in principle be used in the context of forward or inverse dynamics simulations, but in practice this becomes difficult. Forward dynamics simulations would require joint torques—or muscle actuations or excitations in the case of muscle models—as well as torques and forces for the devices as inputs in order to be able to simulate the resulting motions. In the inverse dynamics case, motions, i.e. precise position and velocity profiles, would be required to perform a simulation and compute the required torques, forces, or muscle inputs. However, for complex human motions none of these quantities is precisely known in general. Optimal control is very helpful in the context of motion generation for two reasons:

- It solves the feasibility problem by computing motions and joint torques, etc. that satisfy all the different equality and inequality constraints imposed on the system and the motion task.
- It solves the redundancy problem by determining—from an infinity of different ways to perform a given motion task—the one that is optimal in some sense. Different optimality criteria will be discussed below.

4.1 Multi-Phase Hybrid Optimal Control Problems with Standard Criteria

The task to generate an optimal motion for a model described in Sect. 2 results in the following multi-phase optimal control problem:

$$\min_{x(\cdot),u(\cdot),p,\tau} \sum_{j=1}^{n_{ph}} \left(\int_{\tau_{j-1}}^{\tau_j} \phi_j(x(t),u(t),p)\, dt + \Phi_j(\tau_j, x(\tau_j), p) \right) \tag{38}$$

$$\text{s. t.} \quad \dot{x}(t) = f_j(t,x(t),u(t),p) \quad \text{for} \quad t \in [\tau_{j-1}, \tau_j], \quad j = 1,\ldots,n_{ph},\ \tau_0 = 0, \tau_{n_{ph}} = T \tag{39}$$

$$x(\tau_j^+) = x(\tau_j^-) + J(\tau_j^-, x(\tau_j^-), p) \quad \text{for} \quad j = 1,\ldots,n_{ph} \tag{40}$$

$$g_j(t,x(t),u(t),p) \geq 0 \qquad \text{for} \quad t \in [\tau_{j-1}, \tau_j] \tag{41}$$

$$r_{eq}(x(0),..,x(T),p) = 0 \tag{42}$$

$$r_{ineq}(x(0),..,x(T),p) \geq 0. \tag{43}$$

In these equations, $x(t)$ denotes the vector of state variables, summarizing position and velocity variables, and $u(t)$ is the vector of control variables of the system. In the case of pure rigid body systems (Sects. 2.1 and 2.1), these are the joint torques $\mathscr{M}_i$ as well as the input variables of the medical devices, in the case of musculoskeletal models, these are the muscle inputs, i.e. the muscle activations or excitations. p is the vector of free model parameters. τ is the vector of phase switching times, and the overall time of the motion is $T = \tau_{n_{ph}}$.

Equations (39) and (40) describe the hybrid system dynamics with continuous and discrete motion phases. Here, ordinary differential equations are used for simplicity of presentation; however in reality, we usually face DAE models for most or part of the phases, as described in Sect. 2.1.

Equation (41) summarizes all continuous inequality constraints, which includes simple lower and upper bounds on all variables as well as more complex relations between several variables. In addition, there are coupled and decoupled pointwise equality (42) and inequality constraints (43), e.g. start and end point constraints on the states, phase switching conditions or periodicity constraints.

Equation (38) describes the objective function that is applied to the motion: the first part $\int \phi_j dt$ gives the general form of integral objective functions of Lagrange type while the second part denotes Φ_j Mayer type objective functions that only depend on values at the end of the respective phase. Typical Mayer type objective functions are minimum phase times, minimum total maneuver time, or maximum distance traveled. Typical examples of Lagrange type objective functions include different types of energy or effort minimization, e.g. minimization of mechanical energy or minimization of weighted torques squared, minimization of muscle excitations or activations (to different powers), efficiency maximization, etc. Also some of the stability criteria such as angular momentum minimization or ZMP and capture point related criteria can be formulated as Lagrange type functions. A special form used in context with human movement data is a least squares function minimizing the square between computed (position) trajectories and measured ones. In practice, this is often not a continuous function since measurement points are discrete, but takes the form of a sum over all measurement points, the number of which depends on measurement frequency and phase times:

$$\min_{x(\cdot),u(\cdot),\tau} \sum_{j=1}^{n_{ph}} \sum_{m=1}^{n_M,j} (x'(t_{jm}) - {x_M}'(t_{jm}))^T W (x'(t_{jm}) - {x_M}'(t_{jm})). \tag{44}$$

Here, the superscript $'$ denotes the subset of state variables that are actually measured directly or indirectly (typically the position variables, not the velocities), where x'_M denotes the measured values and x' the corresponding computed variables at all measurement points t_{jm}.

4.2 Multi-Phase Hybrid Optimal Control Problems with Non-standard Criteria Related to Stability

The computation of the objective function and the whole optimal control problem gets, however, much more involved, if asymptotic stability in the sense of Lyapunov is taken into account based on his first method. As described above in Sect. 3, asymptotic stability of a periodic solution of a periodic nonlinear system requires that for all eigenvalues of the monodromy X(T) associated with the periodic solution, we have $|\lambda_i(X(T)| < 1$, i.e. for the spectral radius $\rho := |\lambda_i(X(T)|_{max} < 1$. In the optimal control context, this can be formulated as an objective function that minimizes the spectral radius

$$\min \rho(X(T)). \tag{45}$$

Please note the comments on necessary projections for some cases in Sect. 3 which have to be applied before computing the maximum eigenvalue.

As described above, $X(T)$ is the monodromy or transfer matrix of the periodic solution which depends on sensitivity information of all state end values with respect to perturbations in all initial values. But this first order derivative information is usually not accessible in the objective function of the optimal control problem, compare (38)–(43). We therefore have to reformulate the optimal control problem with augmented dynamics:

$$\min_{x(\cdot),X(\cdot),u(\cdot),p,\tau} \quad \int_0^T \phi(x(t),u(t),p)\,dt + \Phi(T,x(T),X(T),p) \tag{46}$$

$$\text{s. t.} \quad \dot{x}(t) = f_j(t,x(t),u(t),p) \quad \text{for (*)} \tag{47}$$

$$x(\tau_j^+) = x(\tau_j^-) + J(\tau_j^-, x(\tau_j^-), p) \quad \text{for (**)} \tag{48}$$

$$\dot{X}(t) = \frac{\partial f_j}{\partial x}(t,x(t),u(t),p)X(t) \quad \text{for (*) with} \quad X(0) = I \tag{49}$$

$$X(\tau_j^+) = ((f_{j+1}(\tau_j^+) - f_j(\tau_j^-) - J_t - J_x f_j(\tau_j^-)) \cdot \frac{1}{\dot{s}} s_x^T + I + J_x)X(\tau_j^-) \quad \text{for (**)} \tag{50}$$

$$g_j(t,x(t),u(t),p) \geq 0 \qquad \text{for (*)} \tag{51}$$

$$r_{eq}(x(0),..,x(T),p) = 0 \tag{52}$$

$$r_{ineq}(x(0),..,x(T),X(T),p) \geq 0. \tag{53}$$

$$(*) \quad t \in [\tau_{j-1}, \tau_j], \quad j = 1,\ldots,n_{ph},\ \tau_0 = 0, \tau_{n_{ph}} = T$$

$$(**) \quad j = 1,\ldots,n_{ph}$$

In addition to the original hybrid dynamics, this formulation includes the variational differential equation (49) and the corresponding update formula for sensitivities (50), which takes into account that the state dependent phase switching points would move in time in the presence of perturbations.

An interesting alternative to using stability as an objective function (which sometimes does not lead to very natural motions) is to use stability as an inequality constraint, together with another criterion, e.g. related to energy or efficiency. This results in a constraint of the following form

$$\rho(X(T)) \leq c < 1 \tag{54}$$

which again results in the necessity to compute the monodromy matrix $X(T)$ by means of a solution of the variational equation as described above.

The spectral radius criterion may have a problem of ill-conditioning at points of multiple maximum eigenvalues, since the monodromy matrix is a nonsymmetric matrix. The spectral radius becomes non-differentiable, and sometimes even non-Lipschitz at these points, which may occur no matter if the criterion is used as an objective function or as a constraint. We will, however, see below that a solution has been possible anyway.

4.3 Numerical Solution of Optimal Control Problems

For the solution of both types of optimal control problems, we built upon the direct optimal control methods developed by Bock and co-workers (MUSCOD [10, 41]). This code can be applied to mechanical DAEs of the above form, as we showed in [53] and adapted them to handle index-3 DAEs. Optimal control problems can be considered as infinite-dimensional problems (in the sense that $x(t)$ and $u(t)$ are variables in function space), but they can be transformed into finite dimensional problems by means of discretization. The MUSCOD method is based on a direct approach for control discretization using local base functions, such as piecewise constant or linear functions. State parameterization is performed by the multiple shooting approach which splits the entire integration interval into many smaller ones and transforms the original boundary value problem into a set of initial value problems with corresponding continuity and boundary conditions. We use the same grid for both control discretization and state parameterization. These two steps produce a structured non-linear programming problem which can then be solved by an efficient tailored sequential quadratic programming (SQP) algorithm. We would like to point out that a special feature of this multiple shooting approach (as opposed to, e.g. collocation) is that it still includes a simulation of the full problem dynamics on each of the multiple shooting intervals. The accuracy of these simulations can be chosen independently of the multiple shooting and control grid and the discretization is much finer. The simulation is performed simultaneously to the NLP solution using fast and reliable integrators which are also capable to efficiently and accurately compute sensitivity information by internal numerical differentiation (IND [9]).

SQP techniques in general require second-order differentiable functions which pose a theoretical problem with the Lyapunov stability criterion. Our numerical computations in the past have shown that despite this partial violation of theoretical assumptions the optimal control techniques described above work very well, using finite differences for gradient evaluation, even at non-differentiable points; compare Sect. 10.

5 Formulation and Solution of Inverse Optimal Control Problems for Analysis of Motions in Medical Applications

Another problem of interest in the study of human motion—in normal as well as in pathological cases—is the so-called inverse optimal control problem. Assuming that human motion is always optimal or close to optimal, the question here is how the objective function—or typically combination of objective functions—must be chosen to result in a particular motion. In this context, we are not only interested in a qualitative reasoning on contributing factors but in a precise determination of the different elements of the objective function for a given set of real human motions.

The human motion can be experimentally observed and quite precisely measured by different techniques, such as motion capture systems (optical systems or inertial measurement units), force plates for ground reaction forces, EMG measurements for muscle activity, etc. The inverse optimal control problem consists then in determining, from a movement that is (partly) known from measurements, the optimization criterion that has produced this solution. This problem is hard since it consists in solving a parameter estimation problem within an optimal control problem which results naturally in a bilevel formulation (see below) that addresses the parameter identification problem in the upper level and the optimal control problem solution in the lower level.

The term inverse optimal control for the identification of an objective function in an optimal control problem from measurements was first used by Kalman [37] in the context of linear problems. Heuberger [29] discussed inverse optimization in combinatorial problems.

In the mathematical community, the main interest in the field of inverse optimal control problems is to develop solution methods based on a reformulation of the original problem as one-level problem, treating it as a so-called MPEC (mathematical program with equilibrium constraints). Here the lower level optimal control problem is replaced by the corresponding first order optimality or Karush-Kuhn-Tucker (KKT) conditions, which then are formulated as constraints of the higher level parameter estimation problem [45]. A lot of work is performed on the theory of MPECs formulating appropriate optimality conditions and constraint qualifications (e.g., [15, 76]). If this approach is applied in the context of direct optimal control methods (or first-discretize-then-optimize-methods), the optimality conditions must be formulated for the discretized optimal control problem. One example for an applied method of this type is given in [2] which is based on a state discretization by collocation and a solution of the resulting nonlinear programming problem (NLP) by an interior point method, and which is applied to study human arm movement. Hatz et al. [26] proposed an alternative approach in which the KKT conditions are formulated for an optimal control problem discretized by a direct multiple shooting technique and in which the NLP is solved by sequential quadratic programming (SQP). The method is used to investigate walking motions of cerebral palsy patients [25].

An alternative approach that we proposed for nonlinear inverse optimal control problems [57] is to keep the bilevel formulation using a combination of an efficient direct optimal control technique for the lower level and a gradient-free optimization technique for the upper level. The method which is outlined below has been used to identify objective functions of human locomotion trajectory generation, for whole-body human running motions at moderate speeds on a treadmill [58] and for human yoyo-playing [52]. Recently, a new version of this method has been implemented and is currently used to study walking motions in a variety of situations. Liu et al. [44] studied realistic movement generation for character animation by physics-based models and addressed a similar problem: instead of the objective function, which is assumed to be known, they identify unknown model parameters from

motion capture data using a nonlinear inverse optimization technique. In motion studies there are also several authors that combine methods from reinforcement learning, which is a popular approach in robotics with optimization methods to solve problems of the inverse optimal control type (e.g., [16, 43]).

To be able to solve the inverse optimal control problem, we assume that we are able to establish a set of reasonable independent base functions $\Psi_i(t)$ for the objective function which are usually based on expert guesses from the biomechanical or medical field (e.g., a minimum energy or effort, minimum pain in a particular joint or segment, minimum translational or rotational acceleration or jerk, minimum muscle effort, terms related to the perception of the target, etc.). The use of such base functions which have a physical meaning is generally preferable over purely mathematical base functions since they provide the only way to obtain a result that can be interpreted from the application perspective.

The relative contributions of all base functions $\Psi_i(t)$ are expressed by the respective weight factors α_i, which are precisely the unknowns that we aim to determine by the inverse optimal control approach. The inverse optimal control problem is formulated as

$$\min_{\alpha} \sum_{j=1}^{m} ||z^*(t_j;\alpha) - z_M(t_j)||^2 \tag{55}$$

where $z^*(t;\alpha)$ is the solution of

$$\min_{x,u,T} \int_0^T \left[\sum_{i=1}^{n} \alpha_i \Psi_i(x(t),u(t)) \right] dt \tag{56}$$

$$\text{s. t.} \quad \dot{x}(t) = f(t,x(t),u(t)) \tag{57}$$

$$g(t,x(t),u(t)) \geq 0 \tag{58}$$

$$r_{eq}(x(0),..,x(T)) = 0 \tag{59}$$

$$r_{ineq}(x(0),..,x(T)) \geq 0. \tag{60}$$

This problem is a bilevel optimization problem. In the upper level, we aim to minimize the distance between the measured motion and the computed one by optimizing over the vector of weight parameters α. In the lower level, we solve a (forward) optimal control problem for the current iterate of α in order to compute the solution z^* to evaluate the objective function of the higher level problem. The problem formulation shown here only uses single phase problems for simplicity of presentation, however it is no problem to extend this formulation to the type of multi-phase optimal control problems discussed in the previous section. In fact, many of the inerse optimal control applications that we have treated in the past were of this type (e.g., [58]).

In the above formulation, the vector z represents the observation vector of states and possibly control variables with $z = h(x,u)$, where h is the observation function.

In the lower level, the task is to efficiently solve the forward optimal control problem which arises in each iteration of the upper level. For this, we can use the direct boundary value problem approach MUSCOD that has been described in the previous section. In the upper level, the unknown weight factors α have to be determined such that they solve the least squares objective fitting the computational model to measurements. Every function evaluation here requires the solution of the lower level optimal control problem, i.e. the upper level function cannot be expected to satisfy the usual smoothness assumptions. Derivative information of this function could only be generated in a black-box finite difference way. We therefore prefer to apply a derivative-free optimization technique for the upper level, i.e. it only requires function evaluations and no explicit gradient information. The derivative-free optimization code BOBYQA (Bound Optimization BY Quadratic Approximation) by Michael Powell [65] which can also handle bounds on the free parameters performs particularly well in this context.

6 Model-Based Optimization for Physical Assistive Devices for Geriatric Patients

In this and the following sections, we will describe some examples of applications of model-based optimal control in medical and rehabilitation applications. This section summarizes some of our optimization work performed in the still ongoing EU project MOBOT (www.mobot-project.eu). The goal of the MOBOT project in general is to enhance the mobility of elderly people by designing and controlling appropriate robotic physical assistive devices for the support of daily activities such as walking around, standing up, etc. In contrast to the still common passive physical assistive devices (rollators) that are available on the market, the devices developed in this project are inspired by robotic technology and equipped with a variety of sensors to detect the current state of the person as well as the environment, perform predictions of the future and are adaptive since the segments are movable by motors to provide the best postural support. Two different devices are considered in this context:

- an adaptive rollator type device with two handles that the patient can hold onto;
- a device acting more like a human caretaker by actively holding onto the patients waist area and supporting his motions, also referred to as the nurse-type device.

There is a variety of optimization tasks performed by us in MOBOT, e.g.

1. optimization-based motion synergy for geriatric models and different movement tasks;
2. the optimal design of assistive devices (including kinematic and dynamic structure as well as choice of motors, etc.);
3. the identification of the patients' behavior;
4. the optimal online control of the devices' motions.

Here, we cannot give a full overview of all optimization work performed, but focus on the first type of problems and discuss the example of optimizing supported sit to stand (STS) motions. Standing up from a sitting position is a particular challenging type of motion for the class of patients considered and can already pose problems when other types of motions such as walking can still be performed without difficulty. The optimization results presented here provide input to the design of the two different devices and can be used to predict human behavior in the context of control. The models used were set up based on a qualitative evaluation of geriatric motion capture experiments in the project [22]. In these computations, we want to determine how an external physical assistive device would have to act on the subject in order to provide the best possible support for STS motions. The action of the rollator type device is simulated by external forces acting on the hands of the subject. For the nurse type device several alternatives have been studied; the one presented here is simulated by a force acting on the mid-trunk in upward and forward direction as well as two smaller pushing forces below the knee from the front. The advantage of this approach is to use at this point only the human model and not the model of the device itself is that it allows to determine optimal forces and optimal force insertion trajectories from the perspective of the human without introducing any constraints due to an a priori chosen design of the device.

Figure 3a shows the model used for the study of STS motions. It is a symmetric model in the sagittal plane since standard STS motions can be assumed to satisfy these conditions. Left and right half can be assumed to move synchronously such that the corresponding segments of left and right body half can be combined and we can set up a model with eight segments (head, upper trunk, mid-trunk, pelvis and combined segments for the upper arms, lower arms, thighs and shanks, respectively) and 8 DOF. The motion considered here has two phases: a first preparatory phase still sitting on the chair but already moving, and a second phase in which the person has already left the seat and moves upwards to standing configuration. Also the external forces acting on both hands, knees, etc. are combined to one force. The equations of motions take the general form (1)–(8). Liftoff from the chair takes place when the contact force becomes zero, which is a state-dependent switching function, only implicitly defining switching time.

To this model, we impose an objective function of the following form:

$$\min_{x(.),u(.),\tau} \sum_{i=1}^{2} \int_{\tau_{i-1}}^{\tau_i} \left(\sum_{j=1}^{n_{act}} \left(\alpha u_j^2 + \beta |u_j \dot{\phi}_j| \right) + \gamma \dot{\phi}_{abs,head}^2 + \delta \sum_{k=1}^{n_{ext}} u_{ext,k}^2 \right) dt. \tag{61}$$

The elements of this function and their weights have been chosen such that they result in natural movement for the class of patients considered in a qualitative evaluation; unfortunately, the motion capture data collected was not precise enough to allow to perform inverse optimal control computations with it. Since the goal of the devices developed in MOBOT is not to fully support the motion but to provide partial support, we restrict the external vertical support forces to 25 % of the person's weight for the hand force of the rollator and to 50 % for the force at mid-trunk of the nurse-type device.

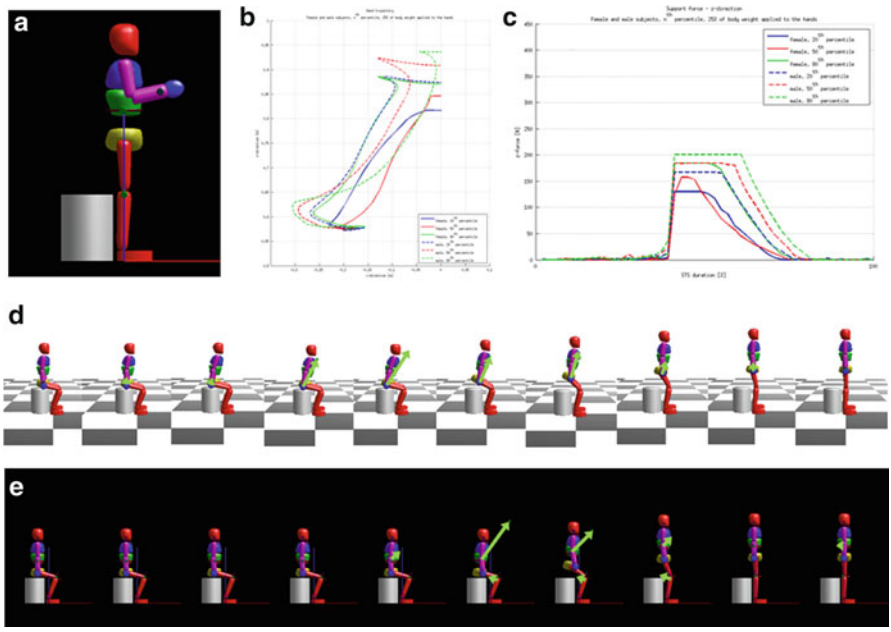

Fig. 3 Sit to stand transfer optimizations with different external forces: Multibody systems model used with 8 DOF in sagittal plane (**a**); Optimal handle trajectories (**b**), optimal profiles for vertical force (**c**) and animation of resulting motion (**d**), all for external forces at hands, mimicking rollator type device; animation of optimal motion for external forces at trunk and knees (**e**), mimicking nurse type device. *Green arrows* in (**d**) and (**e**) show size and direction of external forces

Figure 3 shows exemplary results for the motion of an average (50th percentile) female subject (1.585 m, 64.3 kg) for the rollator type device (d) and the nurse-type device (e). The green arrows depict the orientation and the (relative) value of the external force. Figure 3 shows the corresponding force profiles (c) and force insertion point trajectories (b) for the rollator type device. We have performed such computations for six different sets of anthropomorphic model parameters representing the 20th, 50th, and 80th percentile of male and female geriatric subjects. For these six cases, kinematic and dynamic parameters of all segments are computed using regression formulas specially adapted to the properties of the older population [31]. All results are given in [51].

These results defining the best conditions for humans can then be used as inputs for computations aiming to fix the design of the assistive device. This has been pursued for the rollator in [31] finding the segment lengths joint locations, linear actuator insertion points, actuator characteristics, etc. by again solving a optimal control problem, this time for the dynamic model of the rollator.

7 Optimization and Analysis of Motions with Lower Limb Prostheses

The objective of prostheses is to replace the functionality of a human limb for a variety of desired tasks. For lower limb prostheses, we distinguish transtibial prostheses that replace part of the shank and the foot and transfemural prostheses which replace a large part of the leg and are attached to the body at the level of the thigh. Prostheses for everyday use exist at different levels of actuation, control and sophistication and are designed to provide assistance for standard motions such as walking, standing or standing up. There are also special purpose prostheses for special types of sports, such as running or mountain climbing. In the case of general multi-purpose prostheses the ultimate goal is to achieve a performance and versatility level that is comparable to the one of a healthy human. In contrast, in the case of special purpose prostheses, it is sometimes assumed—but not yet proven—that performance can go even beyond, as we will see below.

Mathematical optimization can be applied in the context of prosthetic motions to match modeled motions to recorded motions or to optimized design and control parameters of the prostheses. It can also be used to analyze the interaction between human and prosthesis. In the control context, real time optimization could be applied to optimize an intelligent prosthesis' reaction during the motion.

The optimization example we are presenting here covers a study of fast running motions with special purpose prosthetic legs. It is discussed in detail in [50]. The case became popular for the South African bilateral transtibial amputee sprinter O. Pistorius who showed such a remarkable performance in the 400 m run that he did not only win the Paralympics but did also get close to the best able-bodied sprinters. A debate started whether he should be excluded from the regular competition or not since his prostheses might provide him with an unfair advantage over able-bodied athletes. The same question appeared again for the German long-jumper M. Rehm who was not allowed to participate in the European championships in 2014 even though he had just won the national competitions.

The special purpose prostheses used for running and jumping are passive torsional carbon-fiber springs that are considerably lighter than human lower legs, can store energy as every spring, but don't have any actuation nor any other means of adjustment. It is not possible to validate the assumption of a potential advantage by means of measurements alone since there are not enough subjects at this high level of performance and the same level of impairment that statistically relevant statements could be inferred. Instead, mathematical modeling, simulation, and optimization can be used to look inside the combined human-prostheses system.

We have performed an optimization study of running motions of a bilateral amputee with the anthropomorphic data corresponding to O. Pistorius as well as of a comparable able-bodied sprinter model for reference. The anthropomorphic model used consists of nine segments (two thighs, shanks, feet and arms as well as a combined trunk-head segment) and describes running in the sagittal plane with 11 DOF. The motion is driven by torques at all internal joints. In the amputee model, the

active torque at the two ankles is replaced by a torque produced by a spring-damper element that has the characteristics of the prosthetic device. We have investigated purely periodic and symmetric running motions which consist of sequences of identical steps such that the model can be reduced to one step consisting of a single support phase, a flight phase, and a touchdown discontinuity. Contact during running only occurs with the ball of the foot. The equations of motion take the form of (1)–(8). We have imposed an average speed of 9 m/s and optimized the integral over the weighted sum of all joint torques squared:

$$\sum_{i=1}^{n_{ph}} \min_{x(\cdot),u(\cdot),\tau} \int_{\tau_{i-1}}^{\tau_i} \sum_{i=1,\ldots,8} (w_i u_i^2)\, dt. \tag{62}$$

The weight factors w_i take into account the maximum torque at each joint. The same problem is solved for the double amputee and the able-bodied model. Image sequences of both results are given in Fig. 4a, b. Part (c) of the figure shows the resulting active joint torques in the able-bodied vs. the amputee sprinter. It is obvious that the required torque effort is much smaller in the amputee sprinter even though the imposed running speed is the same in both cases. Also mechanical energy input is much smaller in the amputee due to the high level of energy storage and release in the spring.

It is, however, too early to conclude that this is a clear indication for an advantage of the double amputee over able-bodied sprinters due to a number of reasons:

- Not only steady-state top speed running, but also start, acceleration and deceleration phase must be considered in the model in order to allow a full evaluation of the course.
- The contact between stump and prosthesis which has been considered as rigid in the computations so far is actually quite flexible sometimes, and should be included as a kind of passive joint with limited range in the motion.
- Other actuators (e.g., knee muscles), and not only the ankle muscles can also be compromised by the amputation and therefore it is not fair in the model to assume comparable joint properties for everything but the ankle.
- The precise objective function for 400 m running and the good performance indicator still have to be determined. In addition to maximum speed and maximum forces, fatigue also might play a role and has to be included in the model. It also might be possible to extend the human models by the relevant muscle models.
- Modeling of pain at the intersection stump and shaft also might be important in order to describe realistic constraints for motions.

Many of the issues discussed are highly individual, and models first have to be developed and then adjusted to subjects by careful parameter estimation. This is all subject to ongoing research, and the question about advantage and disadvantage is still not solved.

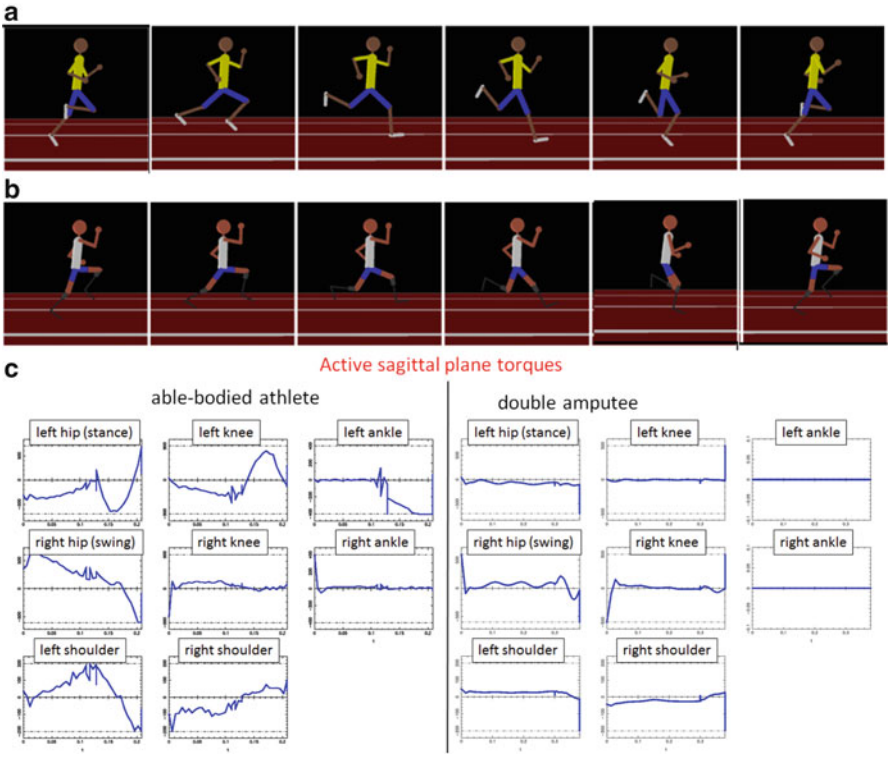

Fig. 4 Some results of model-based optimization for analysis of running with prostheses: optimized motion sequence for able-bodied runner (**a**) and double amputee (**b**), comparison of active joint torques in eight internal joints at hip, knee, ankle, and shoulder (**c**)

8 Model-Based Optimization for the Design of Lower Limb Exoskeletons

The task of orthoses or exoskeletons is not to replace missing limbs, but to provide motion assistance for existing segments of the body by means of an external structure. There are orthoses that only cover one joint, e.g. the knee, and others that cover several or many segments. Orthoses can be purely passive, e.g. consist of elastic and damping material or elements, or they can be actively powered by different types of motors. If an orthosis covers a very large part of the body, e.g. the entire lower or upper extremities or the entire spine, usually the term exoskeleton is used. Exoskeletons have been developed to enhance the existing power of motion of able-bodied people, but also to fully drive motions of paraplegic people, resulting in different required levels of actuation. This section discusses an exoskeleton for the lower extremities as shown in Fig. 1, which was studied in the context of the project HEIKA-EXO within the strategic HEIKA collaboration between the

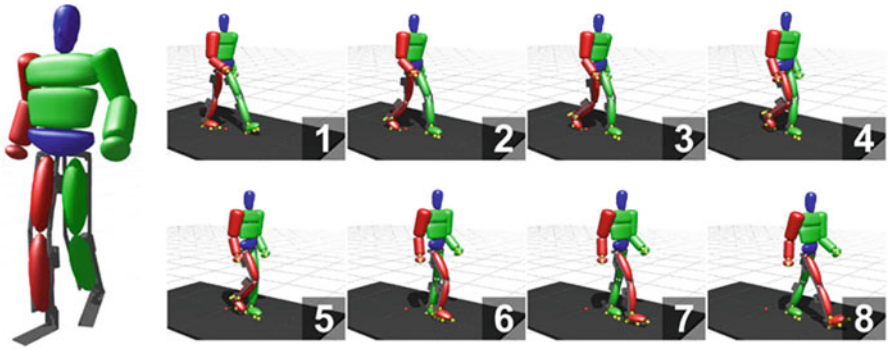

Fig. 5 Combined model of human and lower exoskeleton (*left*, feet not shown); result of least squares optimal control problem for slope down motion

University of Heidelberg and the KIT (www.heika-research.de). This was a small pilot project towards the ultimate goal to develop an exoskeleton as walking support for paraplegics.

When designing an exoskeleton, the challenge is to choose the structure and the powered elements strong enough to support the desired range of motion with the person inside, while at the same time getting not too heavy. Simulations and optimizations that help to make these design choices have to take into account the dynamics of the motions of the combined system human-exoskeleton. In the project, an optimization-based tool called Exo-Opt has been developed to support the design process of exoskeleton, which is described in detail in [39].

For the study presented a 3D whole-body human model with 32 DOF (6 global and 26 at internal joints) has been established as shown in Fig. 5 (left). The exoskeleton covers 12 of these internal joints in the lower extremity. For the first study we have rigidly attached the exoskeleton to the human assuming that they are able to move in perfect conjunction. For later work, an elastic coupling between the two model parts will be considered. The exoskeleton model is kept in a very general form and composed of many different elements with individually adjustable geometry and inertia properties. We have investigated walking motions on level ground as well as up and down different slopes. For all these situations, recordings of an able-bodied person have been taken. The goal was then to determine the joint torques required for the combined human-exoskeleton system to mimic the recorded motion for all position variables $x_{M,l}(l = 1, .., n_{pos})$ by performing a least squares fit to the measurement data in an optimal control formulation

$$\min_{x(\cdot),u(\cdot),p} \sum_{i=1}^{n_{ph}} \left(\sum_{l=1}^{n_{pos}} \sum_{s=1}^{n_M} \alpha_l \left| x_l(t_s) - x_{M,l}(t_s) \right|^2 + \int_{\tau_{i-1}}^{\tau_i} \beta u^T u dt \right). \tag{63}$$

The second term with a small value of β is a regularization term for reducing measurement noise that becomes necessary because of the discretization schemes

of the direct optimal control method and the grid used for the evaluation of the measurements. For the optimal control problem formulation, the mechanical model of the whole walking cycle with all constraints has to be formulated in form of (1)–(8). Walking consists of a sequence of single support and double support phases, but ground contact in walking is more complex than in running since the foot is "rolling" on the ground, and we can distinguish heel only, full foot and toe/ball only contact. This results in four different phases for a walking step with additional discontinuities at touchdown of heel and of toes/ball. This computation has been performed for 15 different combinations of human and exoskeleton masses and inertias and for different slope angles. The results are presented in [39]. Among other information, we have identified:

- the required joint torques as functions of time which help to choose motors and gear types;
- torque-joint angle curves that indicate which motors could be replaced by passive elements like springs;
- structural loads and torques in the constrained DOF as functions of time which serve to support the decision how strong the structure has to be built.

In the design process such a tool can be used in an iterative way: starting with a first estimation for geometry and mass distribution of structure and motors, the tool would be used to compute torques and loads, which then might lead to the necessity to change the structural design and the choice of motors and therefore the dynamic model for which then a new computation would be performed, etc.

The approach presented here is not limited to lower limb exoskeletons but can be used for general exoskeletons and orthoses—as well as for prostheses—if a parameterized dynamic model can be established. This is subject of our ongoing and projected research, e.g. in the context of the new European project SPEXOR (www.spexor.eu) on spinal exoskeletons.

9 Optimization of Functional Electrical Stimulation for Walking Motions of Hemiplegic Patients

FES—sometimes also called neuro-prosthetics—is a technique used in orthopedics and rehabilitation in which the muscles are artificially stimulated by electrical impulses induced locally into the muscle via external or implanted electrodes. FES is often used in patients who have an intact musculo-skeletal system but for whom the neural signal processing does not work, such as different types of paralysis. There is also very promising research on stimulating the nerves in the spinal cord [5] instead of locally in the muscle, but this will not be discussed here. In this section, we discuss optimization work done in the context of the drop foot syndrome of hemiplegic patients that was performed in a collaboration with the LIRMM in Montpellier. The drop foot syndrome is a common problem in hemiplegic patients while walking who are not able to lift the tip of their affected foot such that it is

touching the ground even if a lot of compensatory movement is performed in the hip, and the foot cannot be swung forward. Next to classical, usually rigid, orthoses that keep the ankle joint at a constant right angle, FES is a common treatment for the drop foot syndrome since in this case the stimulation of a single muscle—the tibialis anterior—is usually sufficient [6, 63]. Optimization can be used in this context to address different questions:

- What is the optimal stimulation pattern? Typically simple stimulation patterns, e.g., of trapezoidal shape are used, but these are not necessarily the best possible patterns for the patient. Also a stimulation pattern that leads to a reproduction of a healthy person's motion might not be the optimal choice. Optimization can help to determine optimal patterns with respect to a chosen criterion.
- How can the timing of the stimulation be adjusted to the movement of the healthy leg? What is the good start and end time of stimulation in the gait cycle and which phase shift should be aimed for?
- How can the stimulation process be controlled online? Which sensor information is required for a good state estimation and which information realistically is available? How can nonlinear model predictive control based on optimization methods be used to solve the control problems online?

Here we briefly discuss the first topic, i.e. the generation of motions with optimal stimulation patterns. It should be noted that even though we are interested in a whole-body motion, namely walking, we focus on the part of the motion that can be controlled via the tibialis anterior muscle which is the relative motion in the ankle with respect to the shank. The question is then how to best control the stimulation of the tibialis anterior to lift the foot given a specific or normal motion of the rest of the leg, i.e. given a position and orientation history of the shank during the swing phase of walking. The mechanical multibody system model used in this context therefore does not include the full human body as in the previous three examples, but only the shank and the foot. The torque input at the ankle is replaced by the torque generated by the tibialis anterior muscle which can be described by a Hill type model and the corresponding activation dynamics (see Sect. 2.2). The antagonistic torque is generated by gravity acting on the foot; no other muscle is included in the model so far. The motion considered only covers one phase—the swing phase—starting at lift off and lasting until touch-down of the swing foot. A constraint is formulated to guarantee foot clearance. Different objective functions are studied and compared:

- a least squares fit to a measured motion of a healthy person described by a subset of the state variables, in this case the ankle angle ϕ_{ankle}:

$$\min \sum_{i=1}^{n_M} (\phi_{ankle}(t_i) - \phi_{ankle,ref}(t_i))^2; \tag{64}$$

- a minimization of the metabolic energy consumed which can be expressed in terms of the activation of the tibialis anterior f_{ad}:

$$\min \int_0^T f_{ad}^2 dt; \tag{65}$$

- a minimization of muscle fatigue which can be described by the third power of the activation [1]:

$$\min \int_0^T f_{ad}^3 dt; \tag{66}$$

- criteria related to the muscle excitation ϵ, e.g.

$$\min \int_0^T \epsilon^2 dt. \tag{67}$$

Except for the first criterion for which the time must be fixed, we perform optimizations with free and fixed swing times. Figure 6 shows optimal control results some of which have been presented in [27]. They show clearly that

- for the data studied a perfect fit of the model to measurements is not possible with the tibialis anterior torque alone; an antagonist would be required. It remains to be checked if this is true for general walking motions;

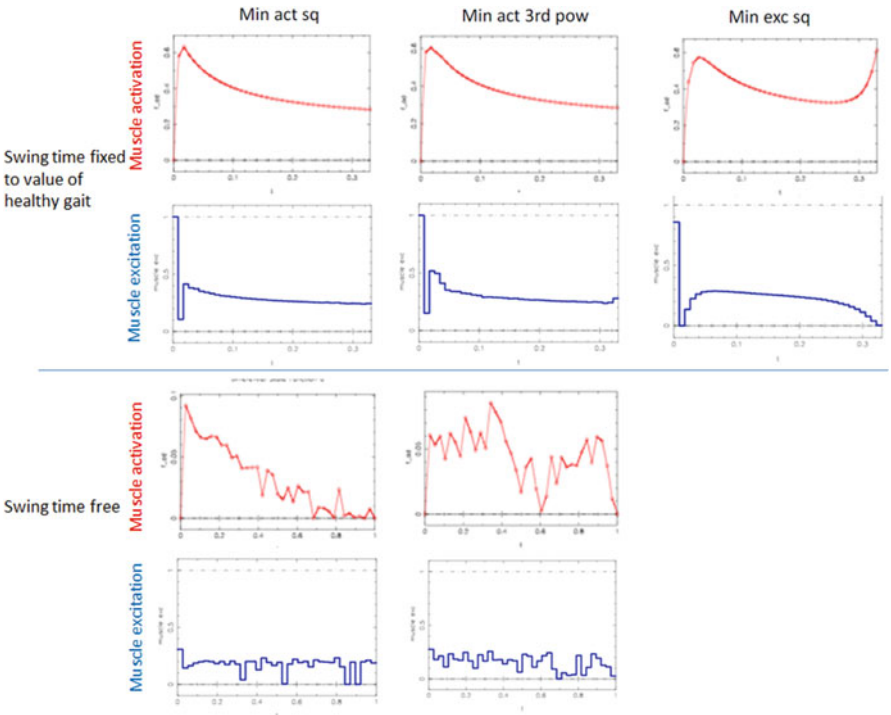

Fig. 6 Optimal stimulation patterns for the tibialis anterior for different objective functions and constraints

- the minimization of energy leads to a much smaller activation level;
- a minimization of fatigue results in activation patterns that show a tendency of having two activation peaks which is related to the patterns observed in healthy humans.

These patterns could be implemented in an open-loop manner on stimulation devices of patients, and their preferences for one or the other input signal could be explored. Currently the focus of this project is on state estimation without drift based on a single IMU at the ankle [7], on the synchronization with the healthy leg and on an implementation of model predictive control algorithms for FES, i.e. a solution of the optimal control problem online, which requires the state estimation as crucial input. This will first be performed in a real time setting in simulation on a computer and on a later stage on the stimulation device on the patient.

10 Stability Studies of Human Walking

This section is different in nature from the previous ones since it does not discuss a specific project, but a topic which is important across all movement studies, in particular in medical and rehabilitation applications, namely the topic of stability. Also for all the examples discussed in the four previous sections, stability is very important, and different criteria presented in Sect. 3 play a role. As mentioned earlier in this paper, the challenge faced in all cases is that in order to control stability by the actions of the medical device, the actions of the human the loop which can only partly be predicted must be compensated by the stability control.

In the case of the assisted STS transfers in the MOBOT project, stability can be defined in terms of static stability asking for the overall center of mass to lie inside the joint polygon of support created by the human and the device. This criterion can be used since the motion is quite slow. A safety margin should be added to compensate for very small dynamic effects. Otherwise also the ZMP criterion discussed in Sect. 3 is applicable.

For other motions with the MOBOT mobility assistance devices like walking, capture point related criteria get interesting. Multiple contacts with the environment have to be considered from the perspective of the human (ground contact and device contact at handles/lower arm/trunk, etc.), and for some of the contacts unilaterality conditions or maximum force conditions (e.g., due to limited wrist force) have to be taken into account. Again, the role of the ZMP of the joint system (human + device) is investigated.

For the more dynamic motions discussed in Sects. 7–9, ZMP and static stability are not very relevant or useful. In the case of the prosthetic running motion, the ZMP criterion cannot be applied at all due to the existence of flight phases and the point contacts in ground contact phase. Also static stability is obviously not defined since there is no polygon of support—just a point (in contact) or nothing at all (in flight). Instead, more dynamic criteria related to the capture point or Lyapunov stability have to be used.

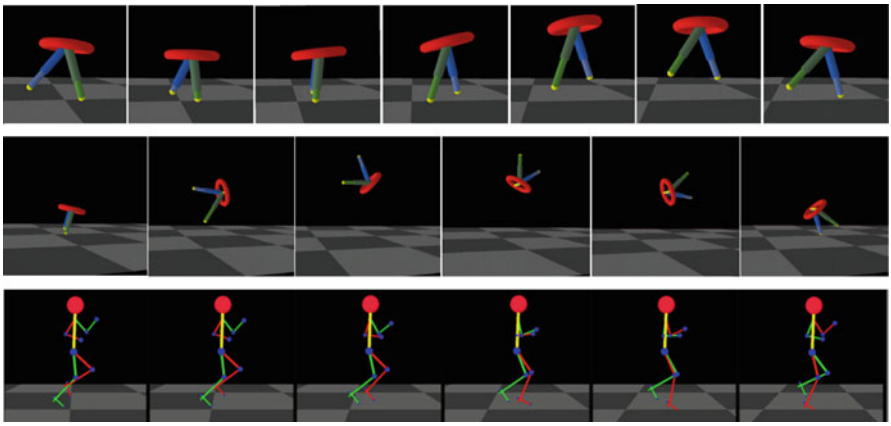

Fig. 7 Results of stability optimization: open-loop stable motions for mono- and bipedal systems. The same approach can be used to explain stability in humans if appropriate models involving relevant feedback are established

We already have investigated the role of the capture point in walking [32] with the goal to extend this study to prosthetic walking. The study showed that subjects step surprisingly close to the capture point.

Lyapunov stability as a truly dynamic stability concept and its mathematical definition has been presented in Sect. 3. In Sect. 4, we have discussed how to formulate Lyapunov stability in the optimal control context as hybrid multi-phase stability optimization problem (Eqs. (46)–(53)). Applying Lyapunov stability criteria to the open-loop optimization to bipedal motion has led to very interesting results for running and different types of jumping and somersaulting [48, 54, 55], see Fig. 7, first two rows. In [49] we have shown open-loop stable running with a planar bipedal model with 11 DOF, similar to the one used in the prostheses studies. For the model we have used with anthropometric data according to de Leva [42]. An animation sequence of the optimal motion is shown in Fig. 7, third row. The solution is stable without any feedback (i.e., the eigenvalues of the monodromy matrix are all smaller than one), but not very robust.

It is important to note that this model may have realistic dynamic parameters for a human, but it is still far away from the human body in terms of explaining human stability control. In this case, we have only investigated the purely open-loop stability of the mechanical system induced by a proper choice of the highly dynamic trajectories and mechanical feedback. In order to explain human stability control also all feedback loops of the human body would have to be included in the model. This includes all relevant muscles (that also exhibit a mechanical kind of self-stability) as well as all neural control loops.

A first step in this direction has been performed in [56] where muscles have been included in a model for juggling motion in order to investigate the role of self-stabilizing properties of muscles. In order to fully explain stability of human whole-body motions based on Lyapunov's first method, it is still a long way to go.

11 Conclusion and Perspectives

In this paper, we have discussed the role of movement primitive in medical applications involving movement studies of healthy and pathological locomotion. From the medical perspective there is a need to better understand human movement and to evaluate how medical and rehabilitation technology should be designed and controlled in order to best support the human. In the introduction, we have listed a number of challenges that arise in this field. For several of these challenges, we also have presented solutions:

- the efficient formulation of personalized whole-body models of human movement with multiple phases;
- some steps towards the formulation of neuromuscular elements in the models;
- general approaches of modeling medical devices and combining them with the human model;
- the formulation of complex movement criteria such as stability which result in non-standard optimal control problems;
- the formulation and solution of optimal control problems for motion generation;
- the identification of situation-specific human optimization criteria from measurements by means of inverse optimal control.

We have demonstrated the usefulness of these methods on some practical examples that we have worked on in interdisciplinary projects, ranging from prostheses and orthoses/exoskeletons over external mobility aids and FES to some general stability studies. All these examples had different requirements with respect to the models and the optimization methods. However, as we have mentioned in several places in this paper, there are still many open challenges and this is still a very active field of research.

We expect the most important research topics in the coming years to be:

- **Subject specific modeling**: Having a subject-specific model of the patient currently under investigation is crucial for obtaining meaningful optimization results. There are several characteristics that have to be adjusted specifically to the subject. Kinematic and dynamic parameters of the multibody dynamics can partly be identified from kinematic measurements but can also be refined by imaging techniques, body material measurements, etc. Also subject-specific parameters for the muscles are required. In addition, for some pathologies such as cerebral palsy also the model topology such as the orientation of joint axes or of muscle origins and insertion points have to be heavily adjusted to the subject. For amputees, a subject-specific modeling of the stump is crucial. A framework for automatically adjusting all these parameters from all types of measurements of a patient would be very desirable.
- **Neuromuscular modeling**: As mentioned in Sect. 2.2, there is still a long way to go towards a good neuromuscular model of the human, but it is crucial for some applications. The field of neuromechanics which tries to bridge the gap between biomechanics and neuroscience has a focus on this topic and research in motor

control will also contribute with findings on learning and adjustment in different feedback loops. Developments in the next years will not result in a complete model but will partly focus on more detailed models of parts of the body, e.g. just an arm or a leg, and partly on whole-body models with reduced muscles and control loops.

- **Soft tissue modeling**: The models discussed in this paper are essentially rigid body models that partly take into account the force generation processes in the muscles but only from the functional perspective, and not their softness property. However, compliance and softness, not only in the muscles but also in all other tissues, in the joints, etc. are considered an essential property of human movement which also is to be copied in technical systems. It is therefore essential to also be able to include continuous mechanics models, in particular finite element models, in the rigid body models to describe soft elements in humans, the devices and in the contacts with each other and the environment. In the next years we will see a lot of development in this area, not only driven by the field of prosthetics and orthotics, but also by the robotics community that is developing robots with soft components.
- **Online optimization**: In this paper, we have essentially focused on optimal control for the task of motion generation and design optimization, which are offline tasks. But of course, optimal control also plays an important role in the context of online motion control which can be tackled by nonlinear model predictive control (NMPC) methods. NMPC plays an important role also in the MOBOT and FES projects, but the details have not been given here for the sake of brevity, and since this is still work in progress. NMPC computes controls for the system online by solving an optimal control problem on a short finite horizon, starting from initial states provided by a state estimator. NMPC methods for problems of the nonlinear multi-phase type that we are facing here are currently under investigation and will have to be further refined and sped up to make them work online on the full system. Ultimately the NMPC algorithms will also have to be implemented on the chips or onboard computers of the medical devices and therefore should be able to work with these limited resources.
- **Reduced models**: Due to the complexity of the combined model of human and device and the limited online computing power, the development of reduced models for use in online control and optimization will be crucial. They should be as sophisticated as possible (so more complex than the standard pendulum or table-cart models currently used in robotics), but still allow all evaluations to be performed in real time.
- **Stability and robustness of motions**: As outlined in Sects. 3 and 10, stability is a crucial topic in healthy and pathological locomotion, but no unique accepted and generally applicable stability criterion exists. In that sense also the robustness of a solution plays a role which in technology usually refers to the size of the basin of attraction of a stable solution of the nonlinear system and not—as in the mathematical definition—to the first order sensitivities. The development of good and reliable stability and robustness measures for truly dynamic motion will also receive significant attention in coming years with a particular focus on measures that can be evaluated online.

- **Optimality criteria for predicting human movement**: In most motion generation and control tasks for technical devices, the human is in the loop and therefore his behavior must be predicted even though it is not fully predictable. As described above, assuming that the movement is always optimal with respect to some criterion is usually a very helpful average guess. The question remains which criterion to apply in which situation. As discussed in Sect. 5, inverse optimal control can do the job, but it presents a lot of work, and in many cases, no appropriate data is available. It would therefore be very helpful to create a database of potential optimality criteria depending on age, general physical condition, pathology, potentially psychological condition, and particular task. Work performed in the KoroiBot project currently goes in that direction for healthy subjects since we study optimality criteria for walking in different terrains. Similar work will have to be performed for patients with different diseases or injuries, or walking with different devices.

Of course, the list of medical and rehabilitation applications given here is not exhaustive. Due to reasons of space, and since we wanted to focus on motion-related problems, we have omitted some very important medical applications in the wider sense, e.g. the entire fields of surgery planning, drug treatment, etc. which also offer a vast terrain for optimal control methods.

Acknowledgements Parts of this research have been supported by the European Union within the European projects MOBOT (GA 600796) and KoroiBot (GA 611909) and the German Excellence Initiative within the third pillar funding of the University of Heidelberg and the HEIKA Research partnership.

References

1. Ackermann, M., van den Bogert, A.J.: Optimality principles for model-based prediction of human gait. J. Biomech. **43**(6), 1055–1060 (2010)
2. Albrecht, S., Passenberg, C., Sobotka, M., Peer, A., Buss, M., Ulbrich, M.: Optimization criteria for human trajectory formation in dynamic virtual environments. In: Haptics: Generating and Perceiving Tangible Sensations. Lecture Notes in Computer Science. Springer, Berlin (2010)
3. Alexander, R.M.: The gaits of bipedal and quadrupedal animals. Int. J. Robot. Res. **3**(2), 49–59 (1984)
4. Alexander, R.M.: Optima for Animals. Princeton University Press, Princeton (1996)
5. Angeli, C., Edgerton, V.R., Gerasimenko, Y., Harkema, S.: Altering spinal cord excitability enables voluntary movements after chronic complete paralysis in humans. Brain: J. Neurol. **137**, 1394–1409 (2014)
6. Azevedo Coste, C., Héliot, R., Pissard-Gibollet, R., Dussaud, P., Andreu, D., Jérôme, F., Laffont, I.: MASEA: Marche Assistée par Stimulation Électrique Adaptative. D'un déclenchement événementiel à un contrôle continu de la stimulation électrique pour la correction du syndrome de pied tombant chez l'hémiplégique. Sciences et Technologie pour le Handicap, Numéro Spécial Handicap et Mouvement (2010)

7. Benoussaad, M., Sijobert, B., Mombaur, K., Azevedo Coste, C.: Robust foot clearance estimation based on the integration of foot-mounted IMU acceleration data. Sensors **16**(1), 12 (2015)
8. Bernstein, N.: The Coordination and Regulation of Movements. Pergamon, Oxford (1967)
9. Bock, H.G.: Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen. In: Bonner Mathematische Schriften, vol. 183. Universität Bonn, Bonn (1987)
10. Bock, H.G., Plitt, K.J.: A multiple shooting algorithm for direct solution of optimal control problems. In: Proceedings of the 9th IFAC World Congress, Budapest, International Federation of Automatic Control, pp. 242–247 (1984)
11. van den Bogert, A.J.: Tutorial: musculoskeletal model for simulation of walking. In: Dynamic Walking Conference 2011, Jena (2011)
12. Coleman, M.J.: A stability study of a three-dimensional passive-dynamic model of human gait. Ph.D. thesis, Cornell University (1998)
13. Cronin, J.: Differential Equations: Introduction and Qualitative Theory. Marcel Dekker, New York (1994)
14. Delp, S., Anderson, F., Arnold, A., Loan, P., Habib, A., John, C., Guendelman, E., Thelan, D.: Opensim: open-source software to create and analyze dynamic simulations of movement. IEEE Trans. Biomed. Eng. **55**, 1940–1950 (2007)
15. Dempe, S., Gadhi, N.: Necessary optimality conditions for bilevel set optimization problems. Glob. Optim. **39**(4), 529–542 (2007)
16. Doerr, A., Ratliff, N., Bohg, J., Toussaint, M., Schaal, S.: Direct loss minimization inverse optimal control. In: Proceedings of Robotics Science and Systems (RSS) (2015)
17. Englsberger, J., Ott, C.: Gangstabilisierung humanoider roboter mittels capture point regelung. Automatisierungstechnik **60**(11), 692–704 (2012)
18. Featherstone, R.: Rigid Body Dynamics Algorithms. Springer, New York (2007)
19. Felis, M.L.: RBDL - Rigid body dynamics library. http://rbdl.bitbucket.org/ (2012–2015)
20. Felis, M.L.: Modeling emotional aspects of human locomotion. Ph.D. thesis, University of Heidelberg (2015)
21. Felis, M.L., Mombaur, K., Berthoz, A.: An optimal control approach to reconstruct human gait dynamics from kinematic data. In: Proceedings of IEEE International Conference on Humanoid Robots (Humanoids 2015) (2015)
22. Fotinea, S.E., Efthimiou, E., Dimou, A.L., Goulas, T., Karioris, P., Peer, A., Maragos, P., Tzafestas, C., Kokkinos, I., Hauer, K., Mombaur, K., Koumpouros, I., Stanzyk, B.: Data acquisition towards defining a multimodal interaction model for human-assistive robot communication. In: Stephanidis, C., Antona, M. (eds.) Universal Access in Human-Computer Interaction Aging and Assistive Environments. Lecture Notes in Computer Science, vol. 8515, pp. 615–626. Springer, Cham (2014)
23. Gopalakrishnan, A., Modenese, L., Phillips, A.T.M.: A novel computational framework for deducing muscle synergies from experimental joint moments. Front. Comput. Neurosci. **8**, 153 (2014)
24. Goswami, A., Espiau, B., Keramane, A.: Limit cycles and their stability and passive bipedal gaits. In: Proceedings of IEEE International Conference on Robotics and Automation, pp. 246–251 (1996)
25. Hatz, K.: Efficient numerical methods for hierarchical dynamic optimization with application to cerebral palsy gait modeling. Ph.D. thesis, University of Heidelberg (2014)
26. Hatz, K., Schlöder, J.P., Bock, H.G.: Estimating parameters in optimal control problems. SIAM J. Sci. Comput. **34**(3), 1707–1728 (2012)
27. Héliot, R., Mombaur, K., Azevedo Coste, C.: Coupling CPG and optimization to generate muscle activation for drop-foot correction online. Modeling, Simulation and Optimization of Bipedal Walking. Springer, New York (2013)
28. Herr, H., Popovic, M.: Angular momentum in human walking. J. Exp. Biol. **211**, 467–81 (2008)

29. Heuberger, C.: Inverse combinatorial optimization: a survey on problems, methods and results. J. Comb. Optim. **8**(3), 329–361 (2004)
30. Hill, A.V.: The heat of shortening and dynamics constants of muscles. Proc. R. Soc. Lond. B **126**(843), 136–195 (1938)
31. Ho Hoang, K.L., Mombaur, K.: Adjustments to de Leva-anthropometric regression data for the changes in body proportions in elderly humans. J. Biomech. **48**(13), 3741–5 (2015)
32. Ho Hoang, K.L., Mombaur, K., Wolf, S.: Investigating capturability in dynamic human locomotion using multi-body dynamics and optimal control. In: Bock, H.G., Hoang, X.P., Rannacher, R., Schlöder, J.P. (eds.) Modeling, Simulation and Optimization of Complex Processes - HPSC 2012, pp. 83–93. Springer, New York (2014)
33. Hof, A.: The 'extrapolated center of mass' concept suggests a simple control of balance in walking. Hum. Mov. Sci. **27**(1), 112–125 (2008)
34. Hsu, J.C., Meyer, A.U.: Modern Control Principles and Applications. McGraw-Hill, New York (1968)
35. Hurmuzlu, Y.: Dynamics of bipedal gait. Part II: Stability analysis of a planar five-link biped. J. Appl. Mech. **60**, 337–343 (1993)
36. Kajita, S., Kanehiro, F., Kaneko, K., Fujiwara, K., Harada, K., Yokoi, K., Hirukawa, H.: Biped walking pattern generation by using preview control of zero-moment point. In: Proceedings of the 2003 IEEE International Conference on Robotics and Automation (2003)
37. Kalman, R.: When is a linear control system optimal? Trans. ASME J. Basic Eng. D **86**(1), 51–60 (1964)
38. Koch, K.H.: Using model-based optimal control for conceptional motion generation for the humanoid robot HRP-2 14 and design investigations for exoskeletons. Ph.D. thesis, University of Heidelberg (2015)
39. Koch, K.H., Mombaur, K.: ExoOpt – a framework for patient centered design optimization of lower limb exoskeletons. In: Proceedings of IEEE International Conference on Rehabilitation Robotics (ICORR) (2015)
40. Koolen, T., De Boer, T., Rebula, J., Goswami, A., Pratt, J.: Capturability-based analysis and control of legged locomotion. Part 1: Theory and application to three simple gait models. Int. J. Robot. Res. **31**(9), 1094–1113 (2012)
41. Leineweber, D.B., Schäfer, A., Bock, H.G., Schlöder, J.P.: An efficient multiple shooting based reduced SQP strategy for large-scale dynamic process optimization. Part II: Software aspects and applications. Comput. Chem. Eng. **27**, 157–174 (2003)
42. de Leva, P.: Adjustments to Zatsiorsky-Seluyanov's segment inertia parameters. J. Biomech. **29**(9), 1223–1230 (1996)
43. Levine, S., Koltun, V.: Guided policy search. In: Proceedings of ICML (2013)
44. Liu, C.K., Hertzmann, A., Popovic, Z.: Learning physics-based motion style with inverse optimization. ACM Trans. Graph. **24**, 1071–1081 (2005)
45. Luo, Z.Q., Pang, J.S., Ralph, D.: Mathematical Programs with Equilibrium Constraints. Cambridge University Press, Cambridge (1996)
46. Majumdar, A., Ahmadi, A.A., Tedrake, R.: Control design along trajectories with sums of squares programming. In: International Conference on Robotics and Automation (ICRA) (2013)
47. McGeer, T.: Passive dynamic walking. Int. J. Robot. Res. **9**, 62–82 (1990)
48. Mombaur, K.D.: Performing open-loop stable flip-flops - an example for stability optimization and robustness analysis of fast periodic motions. Fast Motions in Robotics and Biomechanics - Optimization and Feedback Control. Lecture Notes in Control and Information Science. Springer, Berlin (2006)
49. Mombaur, K.: Using optimization to create self-stable human-like running. Robotica **27**, 321–330 (2009)
50. Mombaur, K.: A mathematical study of sprinting on artificial legs. In: Bock, H.G., Hoang, X.P., Rannacher, R., Schlöder, J.P. (eds.) Modeling, Simulation and Optimization of Complex Processes - HPSC 2012, pp. 157–168. Springer, New York (2014)

51. Mombaur, K., Ho Hoang, K.L.: How to best support sit to stand transfers of geriatric patients: motion optimization under external forces for the design of physical assistive devices (2015, submitted)
52. Mombaur, K., Sreenivasa, M.: Inverse optimal control as a tool to understand human yoyo playing. In: Proceedings of ICNAAM (2010)
53. Mombaur, K.D., Bock, H.G., Schlöder, J.P., Longman, R.W.: Open-loop stable solution of periodic optimal control problems in robotics. J. Appl. Math. Mech. [Z. Angew. Math. Mech.] **85**(7), 499–515 (2005)
54. Mombaur, K.D., Bock, H.G., Schlöder, J.P., Longman, R.W.: Self-stabilizing somersaults. IEEE Trans. Robot. **21**(6), 1148–1157 (2005)
55. Mombaur, K.D., Longman, R.W., Bock, H.G., Schlöder, J.P.: Open-loop stable running. Robotica **23**(01), 21–33 (2005)
56. Mombaur, K.D., Giesl, P., Wagner, H.: Stability optimization of juggling. In: Proceedings of International Conference on High Performance Scientific Computing 2006. Lecture Notes in Scientific Computing. Springer, Hanoi/Vietnam (2008)
57. Mombaur, K., Truong, A., Laumond, J.P.: From human to humanoid locomotion: an inverse optimal control approach. Auton. Robot. **28**(3) (2010) (Published online 31 Dec 2009)
58. Mombaur, K., Olivier, A.H., Crétual, A.: Forward and inverse optimal control of human running. In: Modeling, Simulation and Optimization of Bipedal Walking, vol. 18. Springer, Berlin (2012)
59. Nakamura, Y., Yamane, K., Suzuki, I., Fujita, Y.: Dynamic computation of musculo-skeletal human model based on efficient algorithm for closed kinematic chains. In: Proceedings of the 2nd International Symposium on Adaptive Motion of Animals and Machines (2003)
60. Pandy, M.: Computer modeling and simulation of human movement. Ann. Rev. Biomed. Eng. **3**, 245–273 (2001)
61. Papachristodoulou, A., Prajna, S.: On the construction of Lyapunov functions using the sum of squares decomposition. In: Proceedings of IEEE Conference on Decision and Control (2002)
62. Popovic, M., Englehart, A., Herr, H.: Angular momentum primitives for human walking: biomechanics and control. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (2004)
63. Popovic, M., Popovic, D., Schwirtlich, L., Sinkjaer, T.: Clinical evaluation of functional electrical therapy (FET) in chronic hemiplegic subjects. Neuromodulation **7**(2), 133–140 (2004)
64. Posa, M., Tobenkin, M., Tedrake, R.: Lyapunov analysis of rigid body systems with impacts and friction via sums-of-squares. In: Proceedings of the 16th International Conference on Hybrid Systems: Computation and Control, pp. 63–72. ACM, New York (2013)
65. Powell, M.J.D.: The Bobyqa algorithm for bound constrained optimization without derivatives. Technical Reports 2009/NA06, Department of Applied Mathematics and Theoretical Physics, Cambridge University (2009)
66. Pratt, J., Tedrake, R.: Velocity-based stability margins for fast bipedal walking. In: Fast Motions in Robotics and Biomechanics - Optimization and Feedback Control. Lecture Notes in Control and Information Science. Springer, Berlin (2006)
67. Rosenbaum, D.A.: Human Motor Control. Academic, San Diego (1991)
68. Sardain, P., Bessonnet, G.: Forces acting on a biped robot, center of pressure-zero moment point. IEEE Trans. Syst. Man Cybern. **34**, 630–637 (2004)
69. Sartori, M., Gizzi, L., Lloyd, D., Farina, D.: A musculoskeletal model of human locomotion driven by a low dimensional set of impulsive excitation primitives. Front. Comput. Neurosci. **7**, 1–22 (2013)
70. Sreenivasa, M., Murai, A., Nakamura, Y.: Modeling and identification of the human arm stretch reflex using a realistic spiking neural network and musculoskeletal model. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2013)
71. Steele, K.M., Tresch, M.C., Perreault, E.J.: The number and choice of muscles impact the results of muscle synergy analyses. Front. Comput. Neurosci. **7**, 105 (2013)

72. Stoer, J., Bulirsch, R.: Numerische Mathematik, vol. 2. Springer, Berlin/Heidelberg (1990)
73. Vukobratovic, M., Borovac, B.: Zero-moment-point – thirty five years of its life. Int. J. Humanoid Rob. **1**(1), 157–173 (2004)
74. Wieber, P.B.: Humans toolbox. http://www.inrialpes.fr/bipop/software/humans/ (2007)
75. Winter, D.A.: Biomechanics and Motor Control of Human Movement, 3rd edn. Wiley, New York (2004)
76. Ye, J.: Necessary and sufficient optimality conditions for mathematical programs with equilibrium constraints. J. Math. Anal. Appl. **307**, 350–369 (2005)

Second-Order Optimality Conditions for Broken Extremals and Bang-Bang Controls: Theory and Applications

Nikolai P. Osmolovskii and Helmut Maurer

Abstract We survey the results on no-gap second-order optimality conditions (both necessary and sufficient) in the Calculus of Variations and Optimal Control, that were obtained in the monographs Milyutin and Osmolovskii (Calculus of Variations and Optimal Control. Translations of Mathematical Monographs. American Mathematical Society, Providence, 1998) and Osmolovskii and Maurer (Applications to Regular and Bang-Bang Control: Second-Order Necessary and Sufficient Optimality Conditions in Calculus of Variations and Optimal Control. SIAM Series Design and Control, vol. DC 24. SIAM Publications, Philadelphia, 2012), and discuss their further development. First, we formulate such conditions for broken extremals in the simplest problem of the Calculus of Variations and then, we consider them for discontinuous controls in optimal control problems with endpoint and mixed state-control constraints, considered on a variable time interval. Further, we discuss such conditions for bang-bang controls in optimal control problems, where the control appears linearly in the Pontryagin-Hamilton function with control constraints given in the form of a convex polyhedron. Bang-bang controls induce an optimization problem with respect to the switching times of the control, the so-called Induced Optimization Problem. We show that second-order sufficient condition for the Induced Optimization Problem together with the so-called strict bang-bang property ensures second-order sufficient conditions for the bang-bang control problem. Finally, we discuss optimal control problems with mixed control-state constraints and control appearing linearly. Taking the mixed

N.P. Osmolovskii (✉)
University of Technology and Humanities, ul. Malczewskiego 20a, 26-600 Radom, Poland
Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01-447 Warszawa, Poland
Moscow State University of Civil Engineering, Jaroslavskoe shosse 26, 129337 Moscow, Russia
e-mail: osmolovski@uph.edu.pl

H. Maurer
Institut für Numerische und Angewandte Mathematik, Westfälische Wilhelms–Universität Münster, Einsteinstr. 62, 48149 Münster, Germany
e-mail: maurer@math.uni-muenster.de

J.-B. Hiriart-Urruty et al. (eds.), *Advances in Mathematical Modeling, Optimization and Optimal Control*, Springer Optimization and Its Applications 109,
DOI 10.1007/978-3-319-30785-5_6

constraint as a new control variable we convert such problems to bang-bang control problems. The numerical verification of second-order conditions is illustrated on three examples.

1 Introduction

We survey some main results presented in the recent monograph of the authors [36] (SIAM, 2012) and also some results obtained in the earlier monograph of Milyutin and Osmolovskii [28] (AMS, 1998). We discuss further developments of these results and give various applications.

Our main goal is to present and discuss the no-gap second-order necessary and sufficient conditions in control problems with bang-bang controls. In [28], it was shown how, by using quadratic conditions for the general problem of the Calculus of Variations with regular mixed equality constraint $g(t,x,u) = 0$, one can obtain quadratic (necessary and sufficient) conditions in optimal control problems in which the control variable enters linearly and the control constraint is given in the form of a convex polyhedron. These features were proved in Milyutin and Osmolovskii [28], who first used the property that the set ex U of vertices of a polyhedron U can be described by a nondegenerate relation $g(u) = 0$ on an open set $\mathcal{Q}$ consisting of disjoint open neighborhoods of vertices. This allowed us to develop quadratic necessary conditions for bang-bang controls. Further, in [28] it was shown that a sufficient condition for a minimum on ex U guarantees (in the problem in which the control enters linearly) the minimum on its convexification U. In this way, quadratic sufficient conditions for bang-bang controls were obtained in Osmolovskii and Maurer [36]. This property, which is not discussed in the present paper, constitutes the main link between the second-order optimality conditions for broken extremals in the Calculus of Variations and the second-order optimality conditions for bang-bang controls in optimal control.

The paper is organized as follows. In Sect. 2, we formulate no-gap second-order conditions for broken extremals in the simplest problem of the Calculus of Variations. In Sect. 3, we consider such conditions for discontinuous controls in optimal control problems on a fixed time interval with endpoint constraints of equality and inequality type and mixed state-control constraints of equality type. In Sect. 4, we present an extension of the results of Sect. 3 to problems on a variable time interval. In Sect. 5, we discuss no-gap conditions for bang-bang controls. Bang-bang controls induce an optimization problem with respect to the switching times of the control that we call the *Induced Optimization Problem* (IOP). We have shown in our monograph [36] that the classical second-order sufficient condition for the IOP, together with the so-called strict bang-bang property, ensures second-order sufficient conditions for the bang-bang control problem. We discuss such conditions in Sect. 6.

In the next two sections, the theoretical results are illustrated by numerical examples. Namely, in Sect. 7, we study the optimal control of the chemotherapy of

HIV, when the control-quadratic objective in [18] of L^2-type is replaced by a more realistic L^1-objective. In Sect. 8, we consider time-optimal controls in two models of two-link robots; cf. [36]. Finally, in Sect. 9, we discuss optimal control problems with running mixed control-state constraints and control appearing linearly. Taking the mixed constraint as a new control variable we convert such problem to a bang-bang control problem. We use this transformation to study extremals in the optimal control problem for the Rayleigh equation.

2 Second-Order Optimality Conditions for Broken Extremals in the Simplest Problem in the Calculus of Variations

2.1 The Simplest Problem in the Calculus of Variations

Let a closed interval $[t_0, t_f]$, two points $a, b \in \mathbb{R}^n$, an open set $\mathscr{Q} \subset \mathbb{R}^{2n+1}$, and a function $L : \mathscr{Q} \mapsto \mathbb{R}$ of class C^2 be given. The simplest problem of the Calculus of Variations has the form

$$(SP) \qquad \text{Minimize } \mathscr{J}(x(\cdot)) := \int_{t_0}^{t_f} L(t, x(t), \dot{x}(t))\, dt, \tag{1}$$

$$x(t_0) = a, \quad x(t_f) = b, \quad (t, x(t), \dot{x}(t)) \in \mathscr{Q}. \tag{2}$$

We consider this problem in the space $W^{1,\infty}$ of Lipschitz continuous functions. The last condition in (2) is assumed to hold almost everywhere. A *weak minimum* is defined as a local minimum in the space $W^{1,\infty}$. We say that a function $x \in W^{1,\infty}([t_0, t_f], \mathbb{R}^{d(x)})$ is *admissible* if x satisfies (2) and, moreover, there exists a compact set $\mathscr{C} \subset \mathscr{Q}$ such that $(t, x(t), \dot{x}(t)) \in \mathscr{C}$ a.e. in $[t_0, t_f]$. Set $u := \dot{x}$ and $w = (x, u)$. We call u the *control*.

Let an admissible function $x^0(t)$ be an *extremal* in the sense that it satisfies the *Euler equation*

$$\frac{d}{dt} L_{\dot{x}} = L_x. \tag{3}$$

Here and in the sequel, partial derivatives are denoted by subscripts. Set

$$u^0(t) := \dot{x}^0(t), \quad w^0(t) = (x^0(t), u^0(t)).$$

Let

$$\bar{w}(\cdot) = (\bar{x}(\cdot), \bar{u}(\cdot)) \in \mathscr{W}_2 := W^{1,2} \times L^2,$$

where $W^{1,2}$ is the space of absolutely continuous functions with square integrable derivative and L^2 is the space of square integrable functions. In the space $\mathscr{W}_2$, let us define the subspace

$$\mathscr{K} := \{\bar{w} \in \mathscr{W}_2 \mid \frac{d}{dt}\bar{x}(t) = \bar{u}(t) \text{ a.e., } \bar{x}(t_0) = \bar{x}(t_f) = 0\}$$

and the quadratic form

$$\begin{aligned}\Omega(\bar{w}) &= \int_{t_0}^{t_f} \langle L_{ww}(t, w^0(t))\bar{w}(t), \bar{w}(t)\rangle \, dt \\ &= \int_{t_0}^{t_f} \big(\langle L_{xx}\bar{x}(t), \bar{x}(t)\rangle + 2\langle L_{xu}\bar{u}(t), \bar{x}(t)\rangle + \langle L_{uu}\bar{u}(t), \bar{u}(t)\rangle\big) \, dt.\end{aligned}$$

The following theorem is well known.

Theorem 1. (a) *If the extremal x^0 is a weak minimum, then $\Omega(\bar{w}) \geq 0$ on $\mathscr{K}$.*
(b) *If $\Omega(\bar{w})$ is positive definite on $\mathscr{K}$, then the extremal x^0 is a (strict) weak minimum.*

As is known, the quadratic conditions in Theorem 1 can be tested via the Jacobi conditions or via bounded solutions to an associated Riccati equation.

For a broken extremal, the quadratic form has to be stated in a different way that allows for the formulation of no-gap necessary and sufficient second-order conditions. We will formulate these conditions and discuss their extensions to different classes of optimal control problems, including bang-bang control problems and problems with mixed constraints and control appearing linearly.

2.2 Second-Order Optimality Conditions for Broken Extremals

Let again $x^0(t)$ be an extremal in the simplest problem (1), (2), and let $u^0(t) = \dot{x}^0(t)$ be the corresponding control. Assume now that the control $u^0(t)$ is *piecewise continuous* with one discontinuity point $t_* \in (t_0, t_f)$. Hence, $x^0(t)$ is a *broken extremal* with a corner at t_*. We say that t_* is an *L-point* of the function $u^0(t)$ if there exist $\varepsilon > 0$ and $C > 0$ such that $|u^0(t) - u^0(t_*-)| \leq C|t - t_*|$ for all $t \in (t_* - \varepsilon, t_*)$ and $|u^0(t) - u^0(t_*+)| \leq C|t - t_*|$ for all $t \in (t_*, t_* + \varepsilon)$. Henceforth, we assume that t_* is an *L-point* of the function $u^0(t)$. The following question naturally arises: which quadratic form corresponds to a broken extremal?

Let us change the definition of a weak local minimum as follows. Set $\Theta := \{t_*\}$ and define a notion of a Θ-weak minimum. Assuming additionally that the control $u^0(t)$ is left-continuous at t_*, denote by cl $u^0(\cdot)$ the closure of the graph of $u^0(t)$. Denote by V a neighborhood of the compact set cl $u^0(\cdot)$.

Definition 1. We say that x^0 is a point of a *Θ-weak minimum* (or an *extended weak minimum*) if there exists a neighborhood V of the compact set cl $u^0(\cdot)$ such that $\mathscr{J}(x) \geq \mathscr{J}(x^0)$ for all admissible $x(t)$ such that $u(t) \in V$ a.e., where $u(t) = \dot{x}(t)$.

Clearly, we have the following chain of implications among minima:

$$\text{strong minimum} \quad \Longrightarrow \quad \Theta\text{–weak minimum} \quad \Longrightarrow \quad \text{weak minimum}.$$

Let us formulate optimality conditions for a Θ-weak minimum. To this end, we introduce the Pontryagin function (Hamiltonian)

$$H(t,x,u,\lambda) = \lambda u + L(t,x,u),$$

where λ is a row vector of the dimension n. Defining $\lambda(t) := -L_u(t,x^0(t),u^0(t))$, we have in view of the Euler equation (3):

$$H_u(t,x^0(t),u^0(t),\lambda(t)) = 0, \quad -\dot{\lambda}(t) = H_x(t,x^0(t),u^0(t),\lambda(t)).$$

Denote by $[\lambda]$ the jump of the function $\lambda(t)$ at the point t_*, i.e., $[\lambda] = \lambda^+ - \lambda^-$, where $\lambda^- = \lambda(t_*-)$ and $\lambda^+ = \lambda(t_*+)$. Let $[H]$ stand for the jump of the function $H(t) := H(t,x^0(t),u^0(t),\lambda(t))$ at the same point. The equalities

$$[\lambda] = 0, \quad [H] = 0$$

constitute the *Weierstrass–Erdmann* conditions. They are known as necessary conditions for a strong minimum. However, they are also necessary for the Θ-weak minimum. We add one more necessary condition for the Θ-weak minimum:

$$D(H) := -L_x^+ \dot{x}^{0-} + L_x^- \dot{x}^{0+} - [L_t] \geq 0,$$

where $\dot{x}^{0-} = \dot{x}^0(t_*-)$, $L_x^- = L_x(t_*,x^0(t_*-),u^0(t_*-))$, $[L_t] = L_t^+ - L_t^-$, etc. Clearly,

$$D(H) := \dot{\lambda}^+ \dot{x}^{0-} - \dot{\lambda}^- \dot{x}^{0+} + [\dot{\lambda}_0],$$

where $\lambda_0(t) = -H(t)$ (recall that $\frac{d}{dt}H(t) = H_t(t)$ a.e.). Moreover, it can be shown that $D(H)$ is equal to the negative derivative of the function

$$\Delta H(t) := \lambda(t)[u^0] + L(t,x^0(t),u^0(t_*+)) - L(t,x^0(t),u^0(t_*-))$$

at t_*. The existence of the derivative has been proved and, hence, this derivative can also be calculated as $\frac{d}{dt}(\Delta H)(t_*-)$ or as $\frac{d}{dt}(\Delta H)(t_*+)$.

Now, let us formulate second-order optimality conditions for a Θ-weak minimum. Denote by $P_\Theta W^{1,2}$ the Hilbert space of piecewise continuous functions $\bar{x}(t)$, absolutely continuous on each of the two intervals $[t_0,t_*)$ and $(t_*,t_f]$, and such that their first derivative is square integrable. Any $\bar{x} \in P_\Theta W^{1,2}$ can have a nonzero jump $[\bar{x}] = \bar{x}(t_*+0) - \bar{x}(t_*-0)$ at the point t_*. Let $\bar{t}$ be a numerical parameter. Denote by $Z_2(\Theta)$ the space of triples $\bar{z} = (\bar{t},\bar{x},\bar{u})$ such that $\bar{t} \in \mathbb{R}$, $\bar{x}(\cdot) \in P_\Theta W^{1,2}$, $\bar{u}(\cdot) \in L^2$, i.e.,

$$Z_2(\Theta) = \mathbb{R} \times P_\Theta W^{1,2} \times L^2.$$

In this space, define the quadratic form

$$\Omega_\Theta(\bar z) = D(H)\bar t^2 - 2[L_x]\bar x_{\rm av}\bar t + \int_{t_0}^{t_f} \langle L_{ww}(t,w^0(t))\bar w(t),\bar w(t)\rangle\, dt,$$

where $[L_x]$ is the jump of the function $L_x(t,w^0(t))$ at the point t_*, and

$$\bar x_{\rm av} = \frac{1}{2}\Big(\bar x(t_*-) + \bar x(t_*+)\Big).$$

Set

$$\mathscr{K}_\Theta = \{\bar z \in Z_2(\Theta) \;\mid\; \frac{d}{dt}\bar x(t) = \bar u(t) \;\text{ a.e.}, \quad [\bar x] + [\dot x^0]\bar t = 0, \quad \bar x(t_0) = \bar x(t_f) = 0\,\}.$$

Theorem 2. *(a) If x^0 is a Θ-weak minimum, then $\Omega_\Theta(\bar z) \ge 0$ on $\mathscr{K}_\Theta$. (b) If $\Omega_\Theta(\bar z)$ is positive definite on $\mathscr{K}_\Theta$, then x^0 is a (strict) Θ-weak minimum.*

The proof of this theorem is given in [28]. Let us note that in [28], instead of $\bar t$, we used a numerical parameter $\bar\xi$ such that $\bar t = -\bar\xi$. This remark also applies to the subsequent presentation.

3 Second-Order Optimality Conditions for Discontinuous Controls in the General Problem of the Calculus of Variations on a Fixed Time Interval

3.1 The General Problem in the Calculus of Variations on a Fixed Time Interval

Now consider the following optimal control problem in Mayer form on a fixed time interval $[t_0,t_f]$. It is required to find a pair of functions $w(t) = (x(t),u(t))$, $t \in [t_0,t_f]$, minimizing the functional

$$\min \;\mathscr{J}(w) := J(x(t_0),x(t_f)) \tag{4}$$

subject to the constraints

$$\left.\begin{array}{ll} F(x(t_0),x(t_f)) \le 0, & K(x(t_0),x(t_f)) = 0, \quad (x(t_0),x(t_f))) \in \mathscr{P}, \\ \dot x(t) = f(t,x(t),u(t)), & h(t,x(t),u(t)) = 0, \quad (t,x(t),u(t)) \in \mathscr{Q}, \end{array}\right\} \tag{5}$$

where $\mathscr{P}$ and $\mathscr{Q}$ are open sets, x, u, F, K, f, and h are vector-functions.

We assume that J, F, and K are defined and twice continuously differentiable on $\mathscr{P}$, and f and h are defined and twice continuously differentiable on $\mathscr{Q}$. It is also

assumed that the gradients with respect to the control $h_{iu}(t,x,u)$, $i=1,\dots,d(h)$ are *linearly independent* at each point $(t,x,u)\in\mathscr{Q}$ such that $h(t,x,u)=0$ (the *regularity assumption* for the equality constraint $h(t,x,u)=0$). Here h_i are the components of the vector function h and $d(h)$ is the dimension of this function.

Problem (1), (2) is considered in the space

$$\mathscr{W} := W^{1,1}([t_0,t_f],\mathbb{R}^n)\times L^\infty([t_0,t_f],\mathbb{R}^m),$$

where $n=d(x)$, $m=d(u)$. Define a norm in this space as a sum of the norms:

$$\|w\| := \|x\|_{1,1}+\|u\|_\infty = |x(t_0)|+\int_{t_0}^{t_f}|\dot{x}(t)|\,dt+\operatorname{esssup}_{[t_0,t_f]}|u(t)|.$$

A *weak minimum* is defined as a local minimum in the space $\mathscr{W}$. We say that $w=(x,u)$ is an *admissible pair* if it belongs to $\mathscr{W}$, satisfies the constraints of the problem, and, moreover, there exists a compact set $\mathscr{C}\subset\mathscr{Q}$ such that $(t,x(t),u(t))\in\mathscr{C}$ for a.a. $t\in[t_0,t_f]$.

It is well known that an optimal control problem with a functional in Bolza form,

$$\min\ \mathscr{J}(w) := J(x(t_0),x(t_f))+\int_{t_0}^{t_f} f_0(t,x(t),u(t))\,dt, \tag{6}$$

can be converted to Mayer form by introducing the ODE $\dot{y}=f_0(t,x,u)$, $y(t_0)=0$.

3.2 First-Order Necessary Conditions

Let $w^0=(x^0,u^0)$ be an *admissible pair*. We introduce the *Pontryagin function* (or the *Hamiltonian*)

$$H(t,x,u,\lambda)=\lambda f(t,x,u)$$

and the *augmented Pontryagin function* (or the *augmented Hamiltonian*)

$$H^a(t,x,u,\lambda,\nu)=H(t,x,u,\lambda)+\nu h(t,x,u),$$

where λ and ν are row-vectors of the dimensions $d(x)=n$ and $d(h)$, respectively. For brevity we set

$$x_0=x(t_0),\quad x_f=x(t_f),\quad \eta=(x_0,x_f).$$

Denote by $\mathbb{R}^{n*}$ the space of n-dimensional row-vectors. Define the *endpoint Lagrange function*

$$l(\eta,\alpha_0,\alpha,\beta)=\alpha_0 J(\eta)+\alpha F(\eta)+\beta K(\eta),$$

where $\alpha_0\in\mathbb{R}$, $\alpha\in(\mathbb{R}^{d(F)})^*$, $\beta\in(\mathbb{R}^{d(K)})^*$. Introduce a tuple of *Lagrange multipliers*

$$\mu=(\alpha_0,\alpha,\beta,\lambda(\cdot),\nu(\cdot))$$

such that $\lambda(\cdot):[t_0,t_f]\to\mathbb{R}^{n*}$ is absolutely continuous and $\nu(\cdot):[t_0,t_f]\to(\mathbb{R}^{d(h)})^*$ is measurable and bounded. Denote by Λ_0 the set of the tuples μ satisfying the following conditions at the point w^0:

$$\begin{aligned}&\alpha_0\geq 0,\ \alpha\geq 0,\ \alpha F(\eta^0)=0,\ \alpha_0+\textstyle\sum_{i=1}^{d(F)}\alpha_i+\sum_{j=1}^{d(K)}|\beta_j|=1,\\ &\dot\lambda=-H^a_x,\quad \lambda(t_0)=-l_{x_0},\quad \lambda(t_f)=l_{x_f},\quad H^a_u=0,\end{aligned}$$

where $\eta^0=(x^0(t_0),x^0(t_f))$, the derivatives l_{x_0} and l_{x_f} are at $(\eta^0,\alpha_0,\alpha,\beta)$ and the derivatives H^a_x, H^a_u are at $(t,x^0(t),u^0(t),\lambda(t),\nu(t))$, $t\in[t_0,t_f]$. By α_i and β_j we denote the components of the row vectors α and β, respectively.

Theorem 3. *If w^0 is a weak local minimum, then Λ_0 is nonempty. Moreover, Λ_0 is a finite dimensional compact set, and the projector $(\alpha_0,\alpha,\beta,\lambda(\cdot),\nu(\cdot))\to(\alpha_0,\alpha,\beta)$ is injective on Λ_0.*

The condition $\Lambda_0\neq\emptyset$ is called the *local Pontryagin minimum principle*, or the *Euler–Lagrange equation*. Let M_0 be the set of all $\mu=(\alpha_0,\alpha,\beta,\lambda(\cdot),\nu(\cdot))\in\Lambda_0$ satisfying the *minimum condition* for a.a. $t\in[t_0,t_f]$:

$$H(t,x^0(t),u,\lambda(t))\geq H(t,x^0(t),u^0(t),\lambda(t))\quad\forall\,u\in U(t,x^0(t)),$$

where

$$U(t,x):=\{u\in\mathbb{R}^m\mid(t,x,u)\in\mathscr{Q},\ h(t,x,u)=0\}.$$

The condition $M_0\neq\emptyset$ is called the *(integral) Pontryagin minimum principle*, which is a necessary condition for the so-called Pontryagin minimum.

Definition 2 (A.A. Milyutin). The pair w^0 affords a *Pontryagin minimum* if for any compact set $\mathscr{C}\subset\mathscr{Q}$ there exists $\varepsilon>0$ such that $\mathscr{J}(w)\geq\mathscr{J}(w^0)$ for all admissible pairs $w(t)=\big(x(t),u(t)\big)$ satisfying the conditions

$$\max_{[t_0,t_f]}|x(t)-x^0(t)|<\varepsilon,\quad\int_{t_0}^{t_f}|u(t)-u^0(t)|<\varepsilon,\quad(t,x(t),u(t))\in\mathscr{C}\ \text{ a.e.}$$

3.3 Second-Order Necessary Conditions

Set

$$\mathscr{W}_2 := W^{1,2}([t_0,t_f],\mathbb{R}^n)\times L^2([t_0,t_f],\mathbb{R}^m),$$

Let $\mathscr{K}$ be the set of all $\bar{w}=(\bar{x},\bar{u})\in\mathscr{W}_2$ satisfying the following conditions:

$$\begin{aligned}
&J'(\eta^0)\bar{\eta}\le 0,\quad F_i'(\eta^0)\bar{\eta}\le 0\ \ \forall i\in I_F(\eta^0),\quad K'(\eta^0)\bar{\eta}=0,\\
&\tfrac{d}{dt}\bar{x}(t)=f_w(t,w^0(t))\bar{w}(t),\ \text{for a.a. } t\in[t_0,t_f],\\
&h_w(t,w^0(t))\bar{w}(t)=0,\ \text{for a.a. } t\in[t_0,t_f],
\end{aligned}$$

where $\bar{\eta}=(\bar{x}(t_0),\bar{x}(t_f))$, $I_F(\eta^0):=\{i: F_i(\eta^0)=0\}$ is the set of active indices. Obviously, $\mathscr{K}$ is a convex cone in the Hilbert space $\mathscr{W}_2$. We call it the *critical cone*.

Let us introduce a quadratic form in $\mathscr{W}_2$. For $\mu\in\Lambda_0$ and $\bar{w}=(\bar{x},\bar{u})\in\mathscr{W}_2$, we set

$$\Omega(\mu,\bar{w})=\langle l^{\mu}_{\eta\eta}(\eta^0)\bar{\eta},\bar{\eta}\rangle+\int_{t_0}^{t_f}\langle H^{a\mu}_{ww}(t)\bar{w}(t),\bar{w}(t)\rangle\,dt,$$

where $l^{\mu}_{\eta\eta}(\eta^0)=l_{\eta\eta}(\eta^0,\alpha_0,\alpha,\beta)$, $H^{a\mu}_{ww}(t)=H^{a}_{ww}(t,x^0(t),u^0(t),\lambda(t),\nu(t))$, and $\bar{\eta}=(\bar{x}(t_0),\bar{x}(t_f))$.

Theorem 4. *If w^0 is a weak minimum, then the set Λ_0 is nonempty and*

$$\max_{\mu\in\Lambda_0}\Omega(\mu,\bar{w})\ge 0\quad \textit{for all}\ \ \bar{w}\in\mathscr{K}.$$

The necessary condition for a Pontryagin minimum differs from this condition only by replacing the set Λ_0 by the set M_0.

Theorem 5. *If w^0 is a Pontryagin minimum, then the set M_0 is nonempty and*

$$\max_{\mu\in M_0}\Omega(\mu,\bar{w})\ge 0\quad \textit{for all}\ \ \bar{w}\in\mathscr{K}.$$

We now assume that the control u^0 is a piecewise continuous function on $[t_0,t_f]$ with the set of discontinuity points $\Theta=\{t_1,\dots,t_s\}$, $t_0<t_1<\dots<t_s<t_f$. We also assume that each $t_k\in\Theta$ is an L-point of the function u^0 (see the definition in Sect. 2.2). In this case, the regularity assumption for h implies that, for any $\mu=(\alpha_0,\alpha,\beta,\lambda(\cdot),\nu(\cdot))\in\Lambda_0$, $\nu(t)$ has the same properties as $u^0(t)$: the function $\nu(t)$ is piecewise continuous and each of its point of discontinuity is an L-point which belongs to Θ. By virtue of the adjoint equation $\dot{\lambda}=-H^a_x$, the same is true for the derivative $\dot{\lambda}(t)$ of the adjoint variable λ. Now, the second-order necessary conditions can be refined as follows.

For $\mu \in M_0$, set

$$D^k(H^{a\mu}) = \dot\lambda^{k+}\dot x^{0k-} - \dot\lambda^{k-}\dot x^{0k+} - [H_t^{a\mu}]^k, \tag{7}$$

where $[H_t^{a\mu}]^k$ is the jump of the derivative $H_t^a(t,x^0(t),u^0(t),\lambda(t),\nu(t))$ at the point t_k, and $\dot\lambda^{k-} := \dot\lambda(t_k-)$, $\dot\lambda^{k+} := \dot\lambda(t_k+)$, etc. Note that $H_t^a = -\dot\lambda_0$, where $\lambda_0(t) = -H(t)$, and hence $-[H_t^a]^k = [\dot\lambda_0]^k$. Sometimes we omit the superscript μ in the notation $D^k(H^{a\mu})$.

We can calculate $D^k(H^a)$ using another method. Namely, $D^k(H^a)$ can be calculated as the derivative of the "jump of H^a" at the point t_k. Introduce the function

$$\begin{aligned}(\Delta_k H^a)(t) &= (\Delta_k H)(t) + (\Delta_k(\nu h))(t)\\ &= \lambda(t)\big(f(t,x^0(t),u^{0k+}) - f(t,x^0(t),u^{0k-})\big)\\ &\quad + \big(\nu^{k+}h(t,x^0(t),u^{0k+}) - \nu^{k-}h(t,x^0(t),u^{0k-})\big).\end{aligned}$$

It can be shown that the function $(\Delta_k H^a)(t)$ is continuously differentiable at the point $t_k \in \Theta$, and its derivative at this point coincides with $-D^k(H^a)$. Therefore, we can obtain the value of $D^k(H^a)$ by calculating the left or right limit of the derivatives of the function $(\Delta_k H^a)(t)$ at the point t_k:

$$D^k(H^a) = -\frac{d}{dt}(\Delta_k H^a)(t_k\pm).$$

For any $\mu \in M_0$, it can be shown that $D^k(H^{a\mu}) \ge 0$, $k = 1,\dots,s$. Set

$$Z_2(\Theta) := \mathbb{R}^s \times P_\Theta W^{1,2}([t_0,t_f],\mathbb{R}^n) \times L^2([t_0,t_f],\mathbb{R}^m),$$

where $P_\Theta W^{1,2}([t_0,t_f],\mathbb{R}^n)$ is the Hilbert space of piecewise continuous functions $x(t)$, absolutely continuous on each interval of the set $[t_0,t_f]\setminus\Theta$ such that their first derivatives are square integrable. Define a *quadratic form* in $Z_2(\Theta)$ as follows:

$$\begin{aligned}\Omega_\Theta(\mu,\bar z) &= \Sigma_{k=1}^s\big(D^k(H^{a\mu})\bar t_k^2 + 2[\dot\lambda]^k\bar x_{\rm av}^k\bar t_k\big)\\ &\quad + \langle l^\mu_{\eta\eta}(\eta^0)\bar\eta,\bar\eta\rangle + \textstyle\int_{t_0}^{t_f}\langle H^{a\mu}_{ww}(t)\bar w(t),\bar w(t)\rangle\,dt,\end{aligned}$$

where $\bar z = (\bar\theta,\bar x,\bar u)$, $\bar\theta = (\bar t_1,\dots,\bar t_s)$, $\bar\eta = (\bar x(t_0),\bar x(t_f))$, $\bar x_{\rm av}^k = \frac12(\bar x(t_k-) + \bar x(t_k+))$, $\bar w = (\bar x,\bar u)$. Define the *critical cone* $\mathscr{K}_\Theta$ in the same space by the relations

$$\begin{aligned}&J'(\eta^0)\bar\eta \le 0, \quad F_i'(\eta^0)\bar\eta \le 0\ \ \forall\, i \in I_F(\eta^0), \quad K'(\eta^0)\bar\eta = 0,\\ &\tfrac{d}{dt}\bar x(t) = f_w(t,w^0(t))\bar w(t), \text{ for a.a. } t \in [t_0,t_f],\\ &[\bar x]^k + [\dot x^0]^k\bar t_k = 0, \quad k = 1,\dots,s\\ &h_w(t,w^0(t))\bar w(t) = 0 \quad \text{for a.a. } t \in [t_0,t_f].\end{aligned}$$

Theorem 6. *If w^0 is a Pontryagin minimum, then the following Condition $\mathscr{A}_\Theta$ holds: the set M_0 is nonempty and*

$$\max_{\mu\in M_0} \Omega_\Theta(\mu,\bar z)\ge 0 \quad \textit{for all } \bar z\in\mathscr{K}_\Theta.$$

Let us give another possible representation for the terms $\left(D^k(H^{a\mu})\bar t_k^2+2[\dot\lambda]^k\bar x_{\rm av}^k\bar t_k\right)$ of the quadratic form $\Omega_\Theta(\mu,\bar z)$ on the critical cone $\mathscr{K}$.

Lemma 1. *Let $\mu\in M_0$ and $z=(\bar\theta,\bar x,\bar u)\in\mathscr{K}_\Theta$. Then, for any $k=1,\dots,s$, the following formula holds*

$$D^k(\bar H^\mu)\bar t_k^2+2[\dot\lambda]^k\bar x_{\rm av}^k\bar t_k=[\dot\lambda_0+\dot\lambda\dot x^0]^k\bar t_k^2+2[\dot\lambda\bar x]^k\bar t_k. \tag{8}$$

Proof. Everywhere in this proof we will omit the subscript and superscript k. Taking into account that

$$D(H^a)=\dot\lambda^+\dot x^{0-}-\dot\lambda^-\dot x^{0+}+[\dot\lambda_0],\quad [\bar x]+[\dot x^0]\bar t=0,$$

we obtain

$$\begin{aligned}
&D(H^a)\bar t^2+2[\dot\lambda]\bar x_{av}\bar t=\bar t^2[\dot\lambda_0]+\bar t^2\left(\dot\lambda^+\dot x^{0-}-\dot\lambda^-\dot x^{0+}\right)+2\bar t[\dot\lambda]\bar x_{av}\\
&=\bar t^2[\dot\lambda_0]+\bar t^2\left([\dot\lambda\dot x^0]-\dot\lambda^+[\dot x^0]-\dot\lambda^-[\dot x^0]\right)+2\bar t[\dot\lambda]\bar x_{av}\\
&=\bar t^2\left([\dot\lambda_0]+[\dot\lambda\dot x^0]\right)+\dot\lambda^+[\bar x]\bar t+\dot\lambda^-[\bar x]\bar t+2\bar t[\dot\lambda]\bar x_{av}\\
&=[\dot\lambda_0+\dot\lambda\dot x^0]\bar t^2+\Big(\dot\lambda^+(\bar x^+-\bar x^-)+\dot\lambda^-(\bar x^+-\bar x^-)+(\dot\lambda^+-\dot\lambda^-)(\bar x^-+\bar x^+)\Big)\bar t\\
&=[\dot\lambda_0+\dot\lambda\dot x^0]\bar t^2+2[\dot\lambda\bar x]\bar t.
\end{aligned}$$

3.4 Second-Order Sufficient Conditions

Here, we will formulate sufficient optimality conditions, but only in the case of *discontinuous* control u^0. Let again u^0 be a piecewise continuous function with the set of discontinuity points Θ and let each $t_k\in\Theta$ be an L-point. A natural strengthening of the necessary condition $\mathscr{A}$ in Theorem 6 turned out to be sufficient not only for the Pontryagin minimum, but also for the so-called bounded strong minimum. This type of minimum will be defined below.

Definition 3. The component x_i of the state vector x is called *unessential* if the functions f and h do not depend on x_i and the functions J, F, and K are affine in $x_i(t_0)$ and $x_i(t_f)$. Let $\underline{x}$ denote the vector composed by essential components of vector x.

For instance, the integral functional $\mathscr{J} = \int_{t_0}^{t_f} f_0(t,x,u)\,dt$ can be brought to the endpoint form: $\mathscr{J} = y(t_f) - y(t_0)$, where $\dot{y} = f_0(t,x,u)$. Clearly, y is unessential component.

Definition 4. An admissible pair w^0 affords a *bounded strong minimum* if for any compact set $\mathscr{C} \subset \mathscr{Q}$ there exists $\varepsilon > 0$ such that $\mathscr{J}(w) \geq \mathscr{J}(w^0)$ for all admissible pairs $w(t) = \big(x(t),u(t)\big)$ satisfying the conditions $|x(t_0) - x^0(t_0)| < \varepsilon$, $\max_{[t_0,t_f]} |\underline{x}(t) - \underline{x}^0(t))| < \varepsilon$ and $(t,x(t),u(t)) \in \mathscr{C}$ a.e. on $[t_0,t_f]$.

Definition 5. An admissible pair w^0 affords a *strong minimum* if there exists $\varepsilon > 0$ such that $\mathscr{J}(w) \geq \mathscr{J}(w^0)$ for all admissible pairs $w(t) = \big(x(t),u(t)\big)$ satisfying the conditions $|x(t_0) - x^0(t_0)| < \varepsilon$ and $\max_{[t_0,t_f]} |\underline{x}(t) - \underline{x}^0(t))| < \varepsilon$.

The following assertion follows from the definitions.

Lemma 2. *If there exists a compact set $\mathscr{C} \subset \mathscr{Q}$ such that $\{(t,x,u) \in \mathscr{Q} : h(t,x,u) = 0\} \subset \mathscr{C}$, then the bounded strong minimum is equivalent to the strong minimum.*

Let us formulate sufficient conditions for a bounded strong minimum. For $\mu \in M_0$, we introduce the following conditions of the *strict minimum principle*:

(a) $H(t,x^0(t),u,\lambda(t)) > H(t,x^0(t),u^0(t),\lambda(t))$
for all $t \in [t_0,t_f] \setminus \Theta$, $u \neq u^0(t)$, $u \in U(t,x^0(t))$,

(b) $H(t_k,x^0(t_k),u,\lambda(t_k)) > H^k$
for all $t_k \in \Theta$, $u \in U(t_k,x^0(t_k))$, $u \neq u^0(t_k-)$, $u \neq u^0(t_k+)$, where
$H^k := H(t_k,x^0(t_k),u^0(t_k-),\lambda(t_k)) = H(t_k,x^0(t_k),u^0(t_k+),\lambda(t_k))$.

We denote by M_0^+ the set of all $\mu \in M_0$ satisfying conditions (a) and (b).

For $\mu \in M_0$ we also introduce the *strengthened Legendre-Clebsch conditions*:

(i) for each $t \in [t_0,t_f] \setminus \Theta$ the quadratic form

$$\langle H^a_{uu}(t,x^0(t),u^0(t),\lambda(t),\nu(t))u,u\rangle$$

is positive definite on the subspace of vectors $u \in \mathbb{R}^m$ such that

$$h_u(t,x^0(t),u^0(t))u = 0.$$

(ii) for each $t_k \in \Theta$, the quadratic form

$$\langle \bar{H}_{uu}(t_k,x^0(t_k),u^0(t_k-),\lambda(t_k),\nu(t_k-))u,u\rangle$$

is positive definite on the subspace of vectors $u \in \mathbb{R}^m$ such that

$$h_u(t_k,x^0(t_k),u^0(t_k-))u = 0.$$

(iii) this condition is symmetric to condition (ii) by replacing (t_k-) everywhere by (t_k+).

Note that for each $\mu \in M_0$ the non-strengthened Legendre–Clebsch conditions hold, i.e., the same quadratic forms are *nonnegative* on the corresponding subspaces.

We denote by $\mathrm{Leg}_+(M_0^+)$ the set of all $\mu \in M_0^+$ satisfying the strengthened Legendre–Clebsch conditions (i)–(iii) and also the conditions

(**iv**) $\quad D^k(H^{a\mu}) > 0$ for all $k = 1, \dots, s$.

Let us introduce the functional

$$\bar{\gamma}(\bar{z}) = \langle \bar{\theta}, \bar{\theta} \rangle + \langle \bar{x}(t_0), \bar{x}(t_0) \rangle + \int_{t_0}^{t_f} \langle \bar{u}(t), \bar{u}(t) \rangle \, dt,$$

where $\bar{z} = (\bar{\theta}, \bar{x}, \bar{u})$ and $\bar{\theta} = (\bar{t}_1, \dots, \bar{t}_s)$.

Theorem 7. *For the pair w^0, assume that the following Condition $\mathscr{B}_\Theta$ holds: the set $\mathrm{Leg}_+(M_0^+)$ is nonempty and there exist a nonempty compact set $M \subset \mathrm{Leg}_+(M_0^+)$ and a number $C > 0$ such that*

$$\max_{\mu \in M} \Omega_\Theta(\mu, \bar{z}) \geq C \bar{\gamma}(\bar{z})$$

for all $\bar{z} \in \mathscr{K}$. Then the pair w^0 affords a (strict) bounded strong minimum.

The sufficient condition $\mathscr{B}_\Theta$ guarantees a certain growth condition for the cost which will be presented below. We define now the concept of the *order function* $\Gamma(t,u)$.

Assuming that the function $u^0(t)$ is left-continuous, denote by cl $u^0(\cdot)$ the closure (in $\mathbb{R}^{m+1}$) of its graph. Denote by cl $u^0(t_{k-1}, t_k)$ the closure in $\mathbb{R}^{m+1}$ of the graph of the restriction of $u^0(t)$ to the interval (t_{k-1}, t_k), $k = 1, \dots, s+1$, where $t_{s+1} = t_f$. Then

$$\mathrm{cl}\, u^0(\cdot) = \bigcup_{k=1}^{s+1} \mathrm{cl}\, u^0(t_{k-1}, t_k).$$

Denote by $\mathscr{V}_k$, $k = 1, \dots, s+1$, a system of non-overlapping neighborhoods of the compact sets cl $u^0(t_{k-1}, t_k)$. Let $\mathscr{V} = \bigcup_{k=1}^{s+1} \mathscr{V}_k$.

Definition 6. The function $\Gamma(t,u) : \mathbb{R}^{1+m} \to \mathbb{R}$ is said to be an *order function* if there exist disjoint neighborhoods $\mathscr{V}_k$ of the compact sets cl $u^0(t_{k-1}, t_k)$ such that the following five conditions hold (Fig. 1):

(1) $\Gamma(t,u) = |u - u^0(t)|^2$ if $(t,u) \in \mathscr{V}_k$, $t \in (t_{k-1}, t_k)$, $k = 1, \dots, s+1$;
(2) $\Gamma(t,u) = 2|t - t_k| + |u - u^{0k-}|^2$ if $(t,u) \in \mathscr{V}_k$, $t > t_k$, $k = 1, \dots, s$;
(3) $\Gamma(t,u) = 2|t - t_k| + |u - u^{0k+}|^2$ if $(t,u) \in \mathscr{V}_{k+1}$, $t < t_k$, $k = 1, \dots, s$;
(4) $\Gamma(t,u) > 0$ if $(t,u) \notin \mathscr{V}$;
(5) $\Gamma(t,u)$ is Lipschitz continuous on each compact set in $\mathbb{R}^{1+m}$.

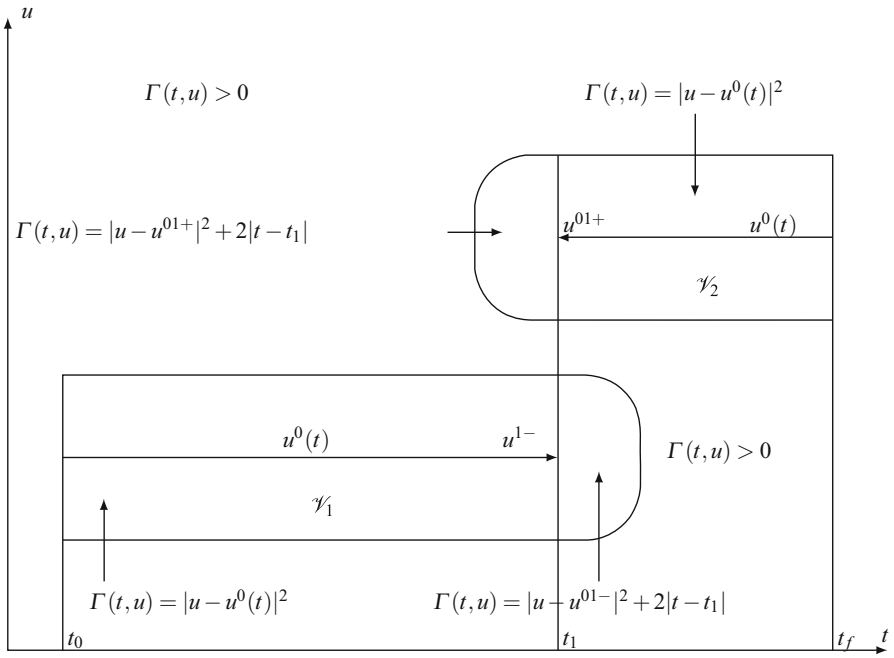

Fig. 1 Illustration of the order function $\Gamma(t,u)$

For $\delta w(t) = (\delta x(t), \delta u(t))$ in $\mathscr{W}$ we set

$$\gamma(\delta w) = \|\delta x\|_\infty^2 + \int_{t_0}^{t_f} \Gamma(t, u^0(t) + \delta u(t))\, dt.$$

We call γ the *higher order*. This higher order corresponds to a typical minimum in the case of discontinuous control, and the order function $\Gamma(t,v)$ corresponds to a typical Hamiltonian in this case.

Note that the order $\int_{t_0}^{t_f} (\Gamma(t, u^0(t) + \delta u(t))\, dt$ is much finer (smaller) than the functional $\int_{t_0}^{t_f} |\delta u(t))|^2\, dt$. On the other hand, it can be proved that, on each compact set (in $\mathbb{R}^m$), the following lower bound holds for $\int_{t_0}^{t_f} \Gamma(t, u^0(t) + \delta u(t))\, dt$:

$$\left(\int_{t_0}^{t_f} |\delta u(t)|\, dt \right)^2 \le C \int_{t_0}^{t_f} \Gamma(t, u^0(t) + \delta u(t))\, dt,$$

where $C > 0$ depends only on the compact set.

Define the *violation function* at the point w^0:

$$V(\delta w) = (J(\eta^0 + \delta\eta) - J(\eta^0))_+ + \sum_{i=1}^{d(F)} F_i(\eta^0 + \delta\eta)_+ + |K(\eta^0 + \delta\eta)| + \|\delta\dot{x} - \delta f\|_1,$$

where $\eta^0 = (x^0(t_0), x^0(t_f))$, $\delta\eta = (\delta x(t_0), \delta x(t_f))$, $\delta f = f(t, w^0 + \delta w) - f(t, w^0)$, $\delta w = (\delta x, \delta u)$, $\|.\|_1$ is the norm in the space L^1 of integrable functions, and $a_+ := \max\{a, 0\}$ for $a \in \mathbb{R}$.

Definition 7. We say that a *bounded strong γ-sufficiency* holds at the point w^0 if there exists $C > 0$ such that for any compact set $\mathscr{C} \subset \mathscr{Q}$ there exists $\varepsilon > 0$ such that the inequality $V(\delta w) \geq C\gamma(\delta w)$ holds for all $\delta w = (\delta x, \delta u) \in \mathscr{W}$ satisfying the conditions

$$\left.\begin{array}{l} |\delta x(t_0)| < \varepsilon,\ \|\underline{\delta x}\|_\infty < \varepsilon, \\ (t, w^0(t) + \delta w(t)) \in \mathscr{C},\ h(t, w^0(t) + \delta w(t)) = 0 \text{ a.e.} \end{array}\right\} \tag{9}$$

Obviously, a bounded strong γ-sufficiency implies a (strict) bounded strong minimum. Moreover, if the point $w^0 + \delta w$ is admissible, then, obviously, $V(\delta w) = (J(w^0 + \delta w) - J(w^0))_+$. Therefore, a bounded strong γ-sufficiency implies the following:

γ-growth condition for the cost: there exists $C > 0$ such that for any compact set $\mathscr{C} \subset \mathscr{Q}$ there exists $\varepsilon > 0$ such that

$$J(w^0 + \delta w) - J(w^0) \geq C\gamma(\delta w)$$

for all $\delta w = (\delta x, \delta u) \in \mathscr{W}$ satisfying (9) and such that $(w^0 + \delta w)$ is an admissible pair.

Theorem 8. *The sufficient condition $\mathscr{B}_\Theta$ in Theorem 7 is equivalent to the bounded strong γ-sufficiency.*

Theorems 6–8 were proved in [31]. Generalizations of these theorems for optimal control problem with regular mixed *inequality* state-control constraints were recently published in [33, 34]. An extension of the results of this section to problems on a variable time interval was obtained in [32].

4 The General Problem in the Calculus of Variations on a Variable Time Interval

4.1 Statement of the Problem

Here, quadratic optimality conditions, both necessary and sufficient, are presented in the following canonical problem on a variable time interval. Let $\mathscr{T}$ denote a process $(x(t), u(t) \mid t \in [t_0, t_f])$, where the state variable $x(\cdot)$ is a Lipschitz continuous function, and the control variable $u(\cdot)$ is a bounded measurable function on a time interval $\Delta = [t_0, t_f]$. The interval Δ is not fixed. For each process $\mathscr{T}$, we denote here by

$$\eta = (t_0, x(t_0), t_f, x(t_f))$$

the vector of the endpoints of time-state variable (t,x). It is required to find $\mathscr{T}$ minimizing the functional

$$\min \mathscr{J}(\mathscr{T}) := J(\eta) \tag{10}$$

subject to the constraints

$$F(\eta) \le 0, \quad K(\eta) = 0, \quad \eta \in \mathscr{P}, \tag{11}$$

$$\dot{x}(t) = f(t,x(t),u(t)), \quad h(t,x(t),u(t)) = 0, \quad (t,x(t),u(t)) \in \mathscr{Q}, \tag{12}$$

where $\mathscr{P}$ and $\mathscr{Q}$ are open sets, x, u, F, K, f, and h are vector-functions.

We assume that the functions J, F, and K are defined and twice continuously differentiable on $\mathscr{P}$, and the functions f and h are defined and twice continuously differentiable on $\mathscr{Q}$. It is also assumed that the gradients with respect to the control $h_{iu}(t,x,u)$, $i = 1,\dots,d(h)$ are linearly independent at each point $(t,x,u) \in \mathscr{Q}$ such that $h(t,x,u) = 0$. Here $d(h)$ is a dimension of the vector h.

4.2 First-Order Necessary Conditions

We say that the function $u(t)$ is *Lipschitz-continuous* if it is piecewise continuous and satisfies the Lipschitz condition on each interval of the continuity. Let

$$\mathscr{T} = (x(t), u(t) \mid t \in [t_0, t_f]) \tag{13}$$

be a fixed admissible process such that the control $u(\cdot)$ is a piecewise Lipschitz-continuous function on the interval Δ with the set of discontinuity points

$$\Theta = \{t_1, \dots, t_s\}, \quad \text{where} \quad t_0 < t_1 < \dots < t_s < t_f.$$

In order to make the notations simpler we do not use here such symbols and indices as zero, hat or asterisk to distinguish the process $\mathscr{T}$ from others.

Let us formulate the first-order necessary condition for optimality of the process $\mathscr{T}$. We introduce the Pontryagin function H (Hamiltonian), the *augmented* Pontryagin function H^a, and the endpoint Lagrange function l as in Sect. 3.2, but remember that now $\eta = (t_0, x_0, t_f, x_f)$, Also we introduce a tuple of Lagrange multipliers

$$\mu = (\alpha_0, \alpha, \beta, \lambda(\cdot), \lambda_0(\cdot), \nu(\cdot)) \tag{14}$$

such that $\lambda(\cdot):\Delta\to(\mathbb{R}^{d(x)})^*$ and $\lambda_0(\cdot):\Delta\to\mathbb{R}^1$ are piecewise smooth functions, continuously differentiable on each interval of the set $\Delta\setminus\Theta$, and $\nu(\cdot):\Delta\to(\mathbb{R}^{d(h)})^*$ is a piecewise continuous function and Lipschitz continuous on each interval of the set $\Delta\setminus\Theta$.

Denote by M_0 the set of the normed tuples μ satisfying the conditions of the *minimum principle* for the process $\mathscr{T}$:

$$\begin{aligned}
&\alpha_0\ge 0,\ \alpha\ge 0,\ \alpha F(\eta)=0,\ \alpha_0+\sum\alpha_i+\sum|\beta_j|=1,\\
&\dot\lambda=-H^a_x,\ \dot\lambda_0=-H^a_t,\ H^a_u=0,\ t\in\Delta\setminus\Theta,\\
&\lambda(t_0)=-l_{x_0},\ \lambda(t_f)=l_{x_f},\ \lambda_0(t_0)=-l_{t_0},\ \lambda_0(t_f)=l_{t_f},\\
&\min_{u\in U(t,x(t))}H(t,x(t),u,\lambda(t))=H(t,x(t),u(t),\lambda(t)),\ t\in\Delta\setminus\Theta,\\
&H(t,x(t),u(t),\lambda(t))+\lambda_0(t)=0,\ t\in\Delta\setminus\Theta,
\end{aligned}\tag{15}$$

where $U(t,x)=\{u\in\mathbb{R}^{d(u)}\mid h(t,x,u)=0,\ (t,x,u)\in\mathscr{Q}\}$. The derivatives l_{x_0} and l_{x_f} are at $(\eta,\alpha_0,\alpha,\beta)$, where $\eta=(t_0,x(t_0),t_f,x(t_f))$, and the derivatives H^a_x, H^a_u, and H^a_t are at $(t,x(t),u(t),\lambda(t),\nu(t))$, where $t\in\Delta\setminus\Theta$. (Condition $H^a_u=0$ follows from other conditions in this definition, and therefore, could be excluded; yet, we need to use it later.)

Let us give the definition of *Pontryagin minimum* in problem (10)–(12) on a variable interval $[t_0,t_f]$.

Definition 8. The process $\mathscr{T}$ affords a *Pontryagin minimum* if for each compact set $\mathscr{C}\subset\mathscr{Q}$ there exists $\varepsilon>0$ such that $\mathscr{J}(\tilde{\mathscr{T}})\ge\mathscr{J}(\mathscr{T})$ for all admissible processes $\tilde{\mathscr{T}}=(\tilde x(t),\tilde u(t)\mid t\in[\tilde t_0,\tilde t_f])$ satisfying the conditions

(a) $|\tilde t_0-t_0|<\varepsilon,\quad |\tilde t_f-t_f|<\varepsilon,$
(b) $\max\limits_{\tilde\Delta\cap\Delta}|\tilde x(t)-x(t)|<\varepsilon,\quad$ where $\tilde\Delta=[\tilde t_0,\tilde t_f],$
(c) $\int\limits_{\tilde\Delta\cap\Delta}|\tilde u(t)-u(t)|\,dt<\varepsilon,$
(d) $(t,\tilde x(t),\tilde u(t))\in\mathscr{C}\quad$ a.e. on $\tilde\Delta$.

The condition $M_0\ne\emptyset$ is equivalent to Pontryagin's minimum principle. It is the first-order necessary condition for Pontryagin minimum for the process $\mathscr{T}$. Thus, the following theorem holds.

Theorem 9. *If the process $\mathscr{T}$ affords a Pontryagin minimum, then the set M_0 is nonempty.*

Assume that the set M_0 is nonempty. Using its definition and the full rank condition for the matrix h_u on the surface $h=0$ one can easily prove the following statement:

Proposition 1. *The set M_0 is a finite-dimensional compact set, and the mapping $\mu\mapsto(\alpha_0,\alpha,\beta)$ is injective on M_0.*

As in Sect. 3, for each $\mu\in M_0, t_k\in\Theta$, we define $D^k(H^{a\mu})$ by relation (7). Then, for each $\mu\in M_0$ the following inequalities hold: $D^k(H^{a\mu})\ge 0,\quad k=1,\dots,s.$

4.3 Second-Order Necessary Conditions

Let us formulate a quadratic necessary condition for a Pontryagin minimum for the process $\mathscr{T}$ as in (13). First, for this process, we introduce a Hilbert space $\mathscr{Z}_2(\Theta)$ and the critical cone $\mathscr{K} \subset \mathscr{Z}_2(\Theta)$. Again, we denote by $P_\Theta W^{1,2}(\Delta, \mathbb{R}^{d(x)})$ the Hilbert space of piecewise continuous functions $\bar{x}(\cdot) : \Delta \to \mathbb{R}^{d(x)}$, absolutely continuous on each interval of the set $\Delta \setminus \Theta$ and such that their first derivative is square integrable. We set

$$\bar{z} = (\bar{t}_0, \bar{t}_f, \bar{\theta}, \bar{x}, \bar{u}),$$

where

$$\bar{t}_0 \in \mathbb{R}^1,\ \bar{t}_f \in \mathbb{R}^1,\ \bar{\theta} = (\bar{t}_1, \dots, \bar{t}_s) \in \mathbb{R}^s,\ \bar{x} \in P_\Theta W^{1,2}(\Delta, \mathbb{R}^{d(x)}),\ \bar{u} \in L^2(\Delta, \mathbb{R}^{d(u)}).$$

Thus,

$$\bar{z} \in \mathscr{Z}_2(\Theta) := \mathbb{R}^2 \times \mathbb{R}^s \times P_\Theta W^{1,2}(\Delta, \mathbb{R}^{d(x)}) \times L^2(\Delta, \mathbb{R}^{d(u)}).$$

Moreover, for given $\bar{z}$ we set

$$\bar{w} = (\bar{x}, \bar{u}),\ \bar{x}_0 = \bar{x}(t_0),\ \bar{x}_f = \bar{x}(t_f), \tag{16}$$

$$\bar{\bar{x}}_0 = \bar{x}(t_0) + \bar{t}_0 \dot{x}(t_0),\ \bar{\bar{x}}_f = \bar{x}(t_f) + \bar{t}_f \dot{x}(t_f),\ \bar{\bar{\eta}} = (\bar{t}_0, \bar{\bar{x}}_0, \bar{t}_f, \bar{\bar{x}}_f). \tag{17}$$

By $I_F(\eta) = \{i \in \{1, \dots, d(F)\} \mid F_i(\eta) = 0\}$ we denote the set of active indices of the constraints $F_i \le 0$. Let $\mathscr{K}_\Theta$ be the set of all $\bar{z} \in \mathscr{Z}_2(\Theta)$ satisfying the following conditions:

$$\begin{aligned}
&J'(\eta)\bar{\bar{\eta}} \le 0, \quad F_i'(\eta)\bar{\bar{\eta}} \le 0\ \forall i \in I_F(\eta), \quad K'(\eta)\bar{\bar{\eta}} = 0,\\
&\tfrac{d}{dt}\bar{x}(t) = f_w(t, w(t))\bar{w}(t), \text{ for a.a. } t \in [t_0, t_f],\\
&[\bar{x}]^k + [\dot{x}]^k \bar{t}_k = 0, \quad k = 1, \dots, s,\\
&h_w(t, w(t))\bar{w}(t) = 0, \text{ for a.a. } t \in [t_0, t_f].
\end{aligned} \tag{18}$$

Clearly, $\mathscr{K}_\Theta$ is a convex cone in the Hilbert space $Z_2(\Theta)$. We call it the *critical cone*. If the interval Δ is fixed, then we set $\eta := (x_0, x_f) = (x(t_0), x(t_f))$, and in the definition of $\mathscr{K}$ we have $\bar{t}_0 = \bar{t}_f = 0$, $\bar{\bar{x}}_0 = \bar{x}_0$, $\bar{\bar{x}}_f = \bar{x}_f$, and $\bar{\bar{\eta}} = \bar{\eta} := (\bar{x}_0, \bar{x}_f)$.

Let us introduce a quadratic form on $\mathscr{Z}_2(\Theta)$. For $\mu \in M_0$ and $\bar{z} \in \mathscr{K}_\Theta$, we set

$$\begin{aligned}
2\Omega_\Theta(\mu, \bar{z}) = {} & \langle l^\mu_{\eta\eta}\bar{\bar{\eta}}, \bar{\bar{\eta}}\rangle + \int_{t_0}^{t_f} \langle H^{a\mu}_{ww}\bar{w}(t), \bar{w}(t)\rangle\, dt + \sum_{k=1}^{s} \Big(D^k(H^{a\mu})\bar{t}_k^2 + 2[\dot{\lambda}]^k \bar{x}^k_{\mathrm{av}} \bar{t}_k \Big)\\
& + \Big(\ddot{\lambda}(t_0)\dot{x}(t_0) + \dot{\lambda}_0(t_0) \Big)\bar{t}_0^2 + 2\dot{\lambda}(t_0)\bar{x}(t_0)\bar{t}_0\\
& - \Big(\ddot{\lambda}(t_f)\dot{x}(t_f) + \dot{\lambda}_0(t_f) \Big)\bar{t}_f^2 - 2\dot{\lambda}(t_f)\bar{x}(t_f)\bar{t}_f,
\end{aligned} \tag{19}$$

where $l^{\mu}_{\eta\eta} = l_{\eta\eta}(\eta, \alpha_0, \alpha, \beta)$, $\quad H^{a\mu}_{ww} = H^{a}_{ww}(t, x(t), u(t), \lambda(t), \nu(t))$. We now formulate the main necessary quadratic condition of Pontryagin minimum in the problem on a variable time interval.

Theorem 10. *If the process $\mathscr{T}$ yields a Pontryagin minimum, then the following Condition $\mathscr{A}_\Theta$ holds: the set M_0 is nonempty and*

$$\max_{\mu \in M_0} \Omega_\Theta(\mu, \bar{z}) \geq 0 \quad \textit{for all } \bar{z} \in \mathscr{K}_\Theta.$$

Using (8), we can represent the quadratic form Ω_Θ on $\mathscr{K}_\Theta$ as follows:

$$\begin{aligned} 2\Omega_\Theta(\mu, \bar{z}) = \langle l^{\mu}_{\eta\eta}\bar{\bar{\eta}}, \bar{\bar{\eta}}\rangle + \int_{t_0}^{t_f} \langle H^{a\mu}_{ww}\bar{w}(t), \bar{w}(t)\rangle\, dt + \sum_{k=1}^{s} \Big([\dot{\lambda}\dot{x} + \dot{\lambda}_0]^k \bar{\tau}_k^2 + 2[\dot{\lambda}\bar{x}]^k \bar{\tau}_k \Big) \\ + \Big(\dot{\lambda}(t_0)\dot{x}(t_0) + \dot{\lambda}_0(t_0) \Big) \bar{\tau}_0^2 + 2\dot{\lambda}(t_0)\bar{x}(t_0)\bar{\tau}_0 \\ - \Big(\dot{\lambda}(t_f)\dot{x}(t_f) + \dot{\lambda}_0(t_f) \Big) \bar{\tau}_f^2 - 2\dot{\lambda}(t_f)\bar{x}(t_f)\bar{\tau}_f. \end{aligned} \tag{20}$$

4.4 Second-Order Sufficient Conditions

Let us give the definition of a bounded strong minimum in problem (10)–(12) on a variable interval $[t_0, t_f]$. Let again $\underline{x}$ denote a vector composed of all essential components of vector x (cf. Definition 3).

Definition 9. The process $\mathscr{T}$ affords *a bounded strong minimum* if for each compact set $\mathscr{C} \subset \mathscr{Q}$ there exists $\varepsilon > 0$ such that $\mathscr{J}(\tilde{\mathscr{T}}) \geq \mathscr{J}(\mathscr{T})$ for all admissible processes $\tilde{\mathscr{T}} = (\tilde{x}(t), \tilde{u}(t) \mid t \in [\tilde{t}_0, \tilde{t}_f])$ satisfying the conditions

(a) $|\tilde{t}_0 - t_0| < \varepsilon, \quad |\tilde{t}_f - t_f| < \varepsilon, \quad |\tilde{x}(\tilde{t}_0) - x(t_0)| < \varepsilon,$
(b) $\max\limits_{\tilde{\Delta} \cap \Delta} |\underline{\tilde{x}}(t) - \underline{x}(t)| < \varepsilon, \quad$ where $\tilde{\Delta} = [\tilde{t}_0, \tilde{t}_f],$
(c) $(t, \tilde{x}(t), \tilde{u}(t)) \in \mathscr{C} \quad$ a.e. on $\tilde{\Delta}$.

The *strict* bounded strong minimum is defined in a similar way, with the nonstrict inequality $\mathscr{J}(\tilde{\mathscr{T}}) \geq \mathscr{J}(\mathscr{T})$ replaced by the strict one and the process $\tilde{\mathscr{T}}$ required to be different from $\mathscr{T}$.

Let us formulate a sufficient optimality Condition $\mathscr{B}_\Theta$, which is a natural strengthening of the necessary Condition $\mathscr{A}_\Theta$. The condition $\mathscr{B}_\Theta$ is sufficient not only for a Pontryagin minimum, but also for a strict bounded strong minimum.

Theorem 11. *For the process $\mathscr{T}$, assume that the following Condition $\mathscr{B}_\Theta$ holds: the set $\mathrm{Leg}_+(M_0^+)$ is nonempty and there exist a nonempty compact set $M \subset \mathrm{Leg}_+(M_0^+)$ and a number $C > 0$ such that*

$$\max_{\mu\in M} \Omega_\Theta(\mu,\bar z) \ge C\bar\gamma(\bar z) \tag{21}$$

for all $\bar z \in \mathscr{K}_\Theta$. *Then the process* $\mathscr{T}$ *affords a strict bounded strong minimum.*

Here the set $\mathrm{Leg}_+(M_0^+)$ has the same definition as in Sect. 3.4.

5 Second-Order Optimality Conditions for Bang-Bang Controls

5.1 Optimal Control Problems with Control Appearing Linearly

Let again $\mathscr{T}$ denote a process $(x(t),u(t)\mid t\in[t_0,t_f])$, where the time interval $\Delta = [t_0,t_f]$ is not fixed. As above, we set

$$\eta = (t_0,x(t_0),t_f,x(t_f)).$$

We will refer to the following control problem (22)–(24) as the *basic control problem*:

$$\text{Minimize } \mathscr{J}(\mathscr{T}) := J(\eta) \tag{22}$$

subject to the constraints

$$F(\eta)\le 0,\quad K(\eta)=0,\quad \eta\in\mathscr{P}, \tag{23}$$

$$\dot x(t) = f(t,x(t)) + g(t,x(t))u(t),\quad u(t)\in U,\quad (t,x(t))\in\mathscr{Q},\quad t_0\le t\le t_f. \tag{24}$$

Here $x\in\mathbb{R}^n$, $u\in\mathbb{R}^m$, F, K, and f are vector functions, g is $n\times m$ matrix function with column vector functions $g_1(t,x,u),\dots,g_m(t,x,u)$, $\mathscr{P}\subset\mathbb{R}^{2n+2}$ and $\mathscr{Q}\subset\mathbb{R}^{n+1}$ are open sets, $U\subset\mathbb{R}^m$ is a convex polyhedron. The functions J, F, and K are assumed to be twice continuously differential on $\mathscr{P}$, and the functions f and g are twice continuously differential on $\mathscr{Q}$.

A process $\mathscr{T} = (x(t),u(t)\mid t\in[t_0,t_1])$ is said to be *admissible* if $x(\cdot)$ is absolutely continuous, $u(\cdot)$ is measurable bounded and the pair of functions $(x(t),u(t))$ on the interval $\Delta=[t_0,t_1]$ with the end-points $\eta = (t_0,x(t_0),t_1,x(t_1))$ satisfies the constraints (23), (24).

Let us give the definition of Pontryagin minimum for the basic problem.

Definition 10. The process $\widehat{\mathscr{T}} = (\hat x(t),\hat u(t)\mid t\in[\hat t_0,\hat t_f])$ affords a *Pontryagin minimum* in the basic problem if there exists $\varepsilon>0$ such that $\mathscr{J}(\mathscr{T})\ge\mathscr{J}(\widehat{\mathscr{T}})$ for all admissible processes $\mathscr{T} = (x(t),u(t)\mid t\in[t_0,t_f])$ satisfying

$$|t_0-\hat{t}_0|<\varepsilon,\quad |t_1-\hat{t}_1|<\varepsilon,\quad \max_{\Delta\cap\widehat{\Delta}}|x(t)-\hat{x}(t)|<\varepsilon,\quad \int\limits_{\Delta\cap\widehat{\Delta}}|u(t)-\hat{u}(t)|\,dt<\varepsilon,$$

where $\Delta=[t_0,t_f]$, $\widehat{\Delta}=[\hat{t}_0,\hat{t}_f]$.

Note that, for a fixed time interval Δ, a Pontryagin minimum corresponds to an L^1-*local minimum* with respect to the control variable.

5.2 *Necessary Optimality Conditions: The Minimum Principle of Pontryagin et al.*

Let $\mathscr{T}=(x(t),u(t)\mid t\in[t_0,t_f])$ be a fixed admissible process such that the control $u(\cdot)$ is a *piecewise constant* function on the interval $\Delta=[t_0,t_f]$. Denote by

$$\Theta=\{t_1,\dots,t_s\},\quad t_0<t_1<\dots<t_s<t_f,$$

the finite set of all discontinuity points (jump points) of the control $u(t)$. Then $\dot{x}(t)$ is a piecewise continuous function whose discontinuity points belong to Θ, and hence $x(t)$ is a piecewise smooth function on Δ.

Let us formulate the *Pontryagin minimum principle*, which is the first-order necessary condition for optimality of the process $\mathscr{T}$. The *Pontryagin function* has the form

$$H(t,x,u,\lambda)=\lambda f(t,x)+\lambda g(t,x)u=\lambda f(t,x)+\sum_{i=1}^{m}\lambda g_i(t,x)u_i, \tag{25}$$

where λ is a row-vector of the dimension $d(\lambda)=d(x)=n$ while x,u,f,F, and K are column-vectors. The factor of the control u in the Pontryagin function is the *switching vector function*, a row vector of dimension $d(u)=m$. Set

$$\begin{aligned}\sigma(t,x,\lambda):&=H_u(t,x,u,\lambda) &&=\lambda g(t,x),\\ \sigma_i(t,x,\lambda):&=H_{u_i}(t,x,u,\lambda) &&=\lambda g_i(t,x),\quad i=1,\dots,m,\\ \sigma_i(t):&=\sigma_i(t,x(t),u(t)).\end{aligned} \tag{26}$$

The *endpoint Lagrange function* is

$$l(\alpha_0,\alpha,\beta,\eta)=\alpha_0 J(\eta)+\alpha F(\eta)+\beta K(\eta),$$

where α and β are row-vectors with $d(\alpha)=d(F)$, $d(\beta)=d(K)$, and α_0 is a number. By

$$\mu=(\alpha_0,\alpha,\beta,\lambda(\cdot),\lambda_0(\cdot))$$

we denote a tuple of Lagrange multipliers such that $\lambda(\cdot) : \Delta \to \mathbb{R}^{n*}$ and $\lambda_0(\cdot) : \Delta \to \mathbb{R}$ are continuous on Δ and continuously differentiable on each interval of the set $\Delta \setminus \Theta$.

Let M_0 be the set of the normed collections μ satisfying the conditions of Minimum Principle for the process $\mathscr{T}$:

$$\alpha_0 \geq 0, \quad \alpha \geq 0, \quad \alpha F(\eta) = 0, \quad \alpha_0 + \sum_{i=1}^{d(F)} \alpha_i + \sum_{j=1}^{d(K)} |\beta_j| = 1, \tag{27}$$

$$\dot{\lambda} = -H_x, \quad \dot{\lambda}_0 = -H_t \quad \forall t \in \Delta \setminus \Theta, \tag{28}$$

$$\lambda(t_0) = -l_{x_0}, \quad \lambda(t_f) = l_{x_f}, \quad \lambda_0(t_0) = -l_{t_0}, \quad \lambda_0(t_f) = l_{t_f}, \tag{29}$$

$$\min_{u \in U} H(t, x(t), u, \lambda(t)) = H(t, x(t), u(t), \lambda(t)) \quad \forall t \in \Delta \setminus \Theta, \tag{30}$$

$$H(t, x(t), u(t), \lambda(t)) + \lambda_0(t) = 0 \quad \forall t \in \Delta \setminus \Theta. \tag{31}$$

The derivatives l_{x_0} and l_{x_f} are taken at the point $(\alpha_0, \alpha, \beta, \eta)$, and the derivatives H_x, H_t are evaluated at the point $(t, x(t), u(t), \lambda(t))$. We use the simple abbreviation (t) for indicating all arguments $(t, x(t), u(t), \lambda(t))$, $t \in \Delta \setminus \Theta$.

Theorem 12. *If the process $\mathscr{T}$ affords a Pontryagin minimum, then the set M_0 is nonempty. The set M_0 is a finite-dimensional compact set and the projector $\mu \mapsto (\alpha_0, \alpha, \beta)$ is injective on M_0.*

In view of this theorem, we can identify each tuple $\mu \in M_0$ with its projection $(\alpha_0, \alpha, \beta)$. In what follows we set $\mu = (\alpha_0, \alpha, \beta)$. For each $\mu \in M_0$ and $t_k \in \Theta$, we define again the quantity $D^k(H^\mu)$. Set

$$(\Delta_k H)(t) = H(t, x(t), u^{k+}, \lambda(t)) - H(t, x(t), u^{k-}, \lambda(t)) = \sigma(t)\,[u]^k. \tag{32}$$

For each $\mu \in M_0$ the following equalities hold:

$$\frac{d}{dt}(\Delta_k H)\big|_{t=t_k-} = \frac{d}{dt}(\Delta_k H)\big|_{t=t_k+}, \quad k = 1, \ldots, s.$$

Consequently, for each $\mu \in M_0$ the function $(\Delta_k H)(t)$ has a derivative at the point $t_k \in \Theta$. Set

$$D^k(H^\mu) = -\frac{d}{dt}(\Delta_k H)(t_k).$$

Then, for each $\mu \in M_0$, the minimum condition (30) implies the inequalities:

$$D^k(H^\mu) \geq 0, \quad k = 1, \ldots, s. \tag{33}$$

As we know, the value $D^k(H^\mu)$ can be written in the form

$$D^k(H^\mu) = -H_x^{k+}H_\lambda^{k-} + H_x^{k-}H_\lambda^{k+} - [H_t]^k = \dot\lambda^{k+}\dot x^{k-} - \dot\lambda^{k-}\dot x^{k+} + [\lambda_0]^k,$$

where H_x^{k-} and H_x^{k+} are the left-hand and the right-hand values of the function $H_x(t) := H_x(t,x(t),u(t),\lambda(t))$ at t_k, respectively, $[H_t]^k$ is a jump of the function $H_t(t)$ at t_k, etc. It also follows from the above representation that we have

$$D^k(H^\mu) = -\dot\sigma(t_k\pm)[u]^k, \tag{34}$$

where the values on the right-hand side agree for the derivative $\dot\sigma(t_k+)$ from the right and the derivative $\dot\sigma(t_k-)$ from the left. In the case of a *scalar* control u, the total derivative $\sigma_t + \sigma_x\dot x + \sigma_\lambda\dot\lambda$ does not contain the control variable explicitly and hence the derivative $\dot\sigma(t)$ is *continuous* at t_k.

Definition 11. For a given extremal process $\mathscr{T} = \{\,(x(t),u(t)) \mid t \in \Delta\,\}$ with a piecewise constant control $u(t)$ we say that $u(t)$ is a *strict bang-bang control* if there exists $\mu = (\alpha_0,\alpha,\beta,\lambda,\lambda_0) \in M_0$ such that

$$\operatorname{Arg\,min}_{u'\in U}\sigma(t)u' = [u(t-),u(t+)]\,, \quad t \in [t_0,t_f] \tag{35}$$

where $[u(t-),u(t+)]$ denotes the line segment spanned by the vectors $u(t-)$ and $u(t+)$ in $\mathbb{R}^{d(u)}$ and $\sigma(t) := \sigma(t,x(t),\lambda(t)) = \lambda(t)g(t,x(t))$.

Note that $[u(t-),u(t+)]$ is a singleton $\{u(t)\}$ at each continuity point of the control $u(t)$ with $u(t)$ being a vertex of the polyhedron U. Only at the points $t_k \in \Theta$ does the line segment $[u^{k-},u^{k+}]$ coincide with an edge of the polyhedron.

It is instructive to evaluate the condition (35) in greater detail when the control set is the *hypercube*

$$U = \prod_{i=1}^{d(u)} [u_{i,min}\,,\, u_{i,max}], \quad u_{i,min} < u_{i,max} \quad (i = 1,\dots,d(u)). \tag{36}$$

Let $S_i = \{t_{k,i}, k = 1,\dots,k_i\}$, $k_i \ge 0$, be the set of switching times of the $i-th$ control component $u_i(t)$ and let $\sigma_i(t) = \lambda(t)g_i(t,x(t))$ be switching function for u_i, $i = 1,\dots,d(u)$. Then the set of all switching times is given by

$$\Theta = \{t_1,\dots,t_s\} = \bigcup_{i=1}^{d(u)} S_i\,,$$

and the condition (35) for a strict bang-bang control requires that

$$\begin{aligned} &(a)\ \sigma_i(t) \neq 0 \quad \forall\, t \notin S_i \quad (i = 1,\dots,d(u)),\\ &(b)\ \text{there is no simultaneous switching of control components } u_i(t),\\ &\quad \text{i.e., } S_i \cap S_j = \emptyset \quad \forall\, i \neq j. \end{aligned} \tag{37}$$

Hence, the i-th control component is determined by the control law

$$u_i(t) = \left\{ \begin{array}{ll} u_{i,\min}, \text{ if } & \sigma_i(t) > 0 \\ u_{i,\max}, \text{ if } & \sigma_i(t) < 0 \end{array} \right\} \quad \forall\, t \in \Delta \setminus S_i. \tag{38}$$

Remark 1. There exist examples, where condition (a) in (37) is slightly violated as $\sigma_i(t_f) = 0$ holds for certain control components u_i; cf. the Rayleigh problem in Sect. 9.2 and the collision avoidance problem in Maurer et al. [27]. In this case, we require in addition that $\dot{\sigma}_i(t_f) \neq 0$ holds to compensate for the condition $\sigma_i(t_f) = 0$. This property is fulfilled for the Rayleigh problem in Sect. 9.2 and the control problem in [27].

5.3 Second-Order Necessary Optimality Conditions

Here, we formulate quadratic necessary optimality conditions for a Pontryagin minimum for a given bang-bang control. (Their strengthening yields quadratic sufficient conditions for a strong minimum.) These quadratic conditions are based on the properties of a quadratic form on the critical cone.

Let again $\mathscr{T} = (x(t), u(t) \mid t \in [t_0, t_f])$ be a fixed admissible process such that the control $u(\cdot)$ is a piecewise constant function on the interval $\Delta = [t_0, t_f]$, and let $\Theta = \{t_1, \ldots, t_s\}$, $t_0 < t_1 < \cdots < t_s < t_f$, be the set of discontinuity points of the control $u(t)$. For the process $\mathscr{T}$, we introduce the space $\mathscr{Z}(\Theta)$ and the *critical cone* $\mathscr{K}_\Theta \subset \mathscr{Z}(\Theta)$ as follows. Denote by $P_\Theta C^1(\Delta, \mathbb{R}^{d(x)})$ the space of piecewise continuous functions $\bar{x}(\cdot) : \Delta \to \mathbb{R}^{d(x)}$ that are continuously differentiable on each interval of the set $\Delta \setminus \Theta$. For each $\bar{x} \in P_\Theta C^1(\Delta, \mathbb{R}^{d(x)})$ and for $t_k \in \Theta$ we set $\bar{x}^{k-} = \bar{x}(t_k-)$, $\bar{x}^{k+} = \bar{x}(t_k+)$ and $[\bar{x}]^k = \bar{x}^{k+} - \bar{x}^{k-}$. Now set

$$\bar{z} = (\bar{t}_0, \bar{t}_f, \bar{\theta}, \bar{x}),$$

where $\bar{t}_0, \bar{t}_f \in \mathbb{R}^1$, $\bar{\theta} = (\bar{t}_1, \ldots, \bar{t}_s) \in \mathbb{R}^s$, $\bar{x} \in P_\Theta C^1(\Delta, \mathbb{R}^{d(x)})$. Thus,

$$\bar{z} \in \mathscr{Z}(\Theta) := \mathbb{R}^2 \times \mathbb{R}^s \times P_\Theta C^1(\Delta, \mathbb{R}^{d(x)}).$$

For each $\bar{z}$ we set

$$\bar{\bar{x}}_0 = \bar{x}(t_0) + \bar{t}_0 \dot{x}(t_0), \quad \bar{\bar{x}}_f = \bar{x}(t_f) + \bar{t}_f \dot{x}(t_f), \quad \bar{\bar{\eta}} = \big(\bar{t}_0, \bar{\bar{x}}_0, \bar{t}_f, \bar{\bar{x}}_f\big). \tag{39}$$

The vector $\bar{\bar{\eta}}$ is considered as a column vector. Note that $\bar{t}_0 = 0$, respectively, $\bar{t}_f = 0$ for *fixed* initial time t_0, respectively, final time t_f. Let

$$I_F(\eta) = \{i \in \{1, \ldots, d(F)\} \mid F_i(\eta) = 0\}$$

be the set of indices of all active endpoint inequalities $F_i \le 0$ at the point $\eta = (t_0, x(t_0), t_f, x(t_f))$. Denote by $\mathscr{K}_\Theta$ the set of all $\bar{z} \in \mathscr{Z}(\Theta)$ satisfying the following conditions:

$$J'(\eta)\bar{\eta} \le 0, \quad F_i'(\eta)\bar{\eta} \le 0 \,\forall i \in I_F(\eta), \quad K'(\eta)\bar{\eta} = 0, \tag{40}$$

$$\frac{d}{dt}\bar{x}(t) = (f_x(t,x(t)) + g_x(t,x(t))u(t))\bar{x}(t), \tag{41}$$

$$[\bar{x}]^k + [\dot{x}]^k \bar{t}_k = 0, \quad k = 1, \dots, s. \tag{42}$$

It is obvious that $\mathscr{K}_\Theta$ is a convex, finite-dimensional, and finite-faced cone in the space $\mathscr{Z}(\Theta)$. We call it *the critical cone*. Each element $\bar{z} \in \mathscr{K}_\Theta$ is uniquely defined by the numbers $\bar{t}_0, \bar{t}_f$, the vector $\bar{\theta}$ and the initial value $\bar{x}(t_0)$ of the function $\bar{x}(t)$. Two important properties of the critical cone are formulated in the next two propositions.

Proposition 2. *For any $\mu \in M_0$ and $\bar{z} \in \mathscr{K}_\Theta$, we have*

$$\alpha_0 J'(\eta)\bar{\eta} = 0, \quad \alpha_i F_i'(\eta)\bar{\eta} = 0 \quad \forall i \in I_F(\eta).$$

Proposition 3. *Suppose that there exist $\mu \in M_0$ with $\alpha_0 > 0$. Then adding the equalities $\alpha_i F_i'(\eta)\bar{\eta} = 0 \;\forall i \in I_F(\eta)$ to the system (40)–(42) defining $\mathscr{K}_\Theta$, one can omit the inequality $J'(\eta)\bar{\eta} \le 0$ in that system without affecting $\mathscr{K}_\Theta$.*

Thus, $\mathscr{K}_\Theta$ is defined by conditions (41), (42) and by the condition $\bar{\eta} \in \mathscr{K}_\Theta^e$, where $\mathscr{K}_\Theta^e$ is the cone in $\mathbb{R}^{2d(x)+2}$ given by (40). But if there exists $\mu \in M_0$ with $\alpha_0 > 0$, then we can put

$$\mathscr{K}_\Theta^e = \{\bar{\eta} \in \mathbb{R}^{d(x)+2} \mid F_i'(\eta)\bar{\eta} \le 0,\ \alpha_i F_i'(\eta)\bar{\eta} = 0 \,\forall i \in I_F(\eta),\ K'(\eta)\bar{\eta} = 0\}. \tag{43}$$

If, in addition, $\alpha_i > 0$ holds for all $i \in I_F(\eta)$, then $\mathscr{K}_\Theta^e$ is a subspace in $\mathbb{R}^{d(x)+2}$.

Let us introduce a quadratic form on the critical cone $\mathscr{K}_\Theta$ defined by the conditions (40)–(42). For each $\mu \in M_0$ and $\bar{z} \in \mathscr{K}_\Theta$ we set

$$\begin{aligned} 2\Omega_\Theta(\mu, \bar{z}) = {} & \langle l^\mu_{\eta\eta}\bar{\eta}, \bar{\eta}\rangle + \int_{t_0}^{t_f} \langle H^\mu_{xx}\bar{x}(t), \bar{x}(t)\rangle\, dt + \sum_{k=1}^{s} \left(D^k(H^\mu)\bar{t}_k^2 + 2[\dot{\lambda}]^k \bar{x}^k_{\mathrm{av}} \bar{t}_k \right) \\ & + \left(\dot{\lambda}(t_0)\dot{x}(t_0) + \dot{\lambda}_0(t_0) \right)\bar{t}_0^2 + 2\dot{\lambda}(t_0)\bar{x}(t_0)\bar{t}_0 \\ & - \left(\dot{\lambda}(t_f)\dot{x}(t_f) + \dot{\lambda}_0(t_f) \right)\bar{t}_f^2 - 2\dot{\lambda}(t_f)\bar{x}(t_f)\bar{t}_f, \end{aligned} \tag{44}$$

where $l^\mu_{\eta\eta} = l_{\eta\eta}(\eta, \alpha_0, \alpha, \beta)$, $\quad H^\mu_{xx} = H_{xx}(t, x(t), u(t), \lambda(t))$ and $\bar{\eta}$ was defined in (39). Note that for a problem on a fixed time interval $[t_0, t_f]$ we have $\bar{t}_0 = \bar{t}_f = 0$. The following theorem gives the main second-order necessary condition of optimality.

Theorem 13. *If the process $\mathscr{T}$ affords a Pontryagin minimum, then the following Condition $\mathscr{A}_\Theta$ holds: the set M_0 is nonempty and* $\max_{\mu\in M_0} \Omega_\Theta(\mu,\bar z) \ge 0$ *for all* $\bar z \in \mathscr{K}_\Theta$.

Using (8), we can also represent the quadratic form Ω_Θ as follows:

$$2\Omega_\Theta(\mu,\bar z) = \langle l^\mu_{\eta\eta}\bar\eta,\bar\eta\rangle + \int_{t_0}^{t_f}\langle H^\mu_{xx}\bar x(t),\bar x(t)\rangle\,dt + \sum_{k=1}^{s}\Big([\dot\lambda\dot x+\dot\lambda_0]^k\bar t_k^2 + 2[\dot\lambda\bar x]^k\bar t_k\Big)$$
$$+\Big(\dot\lambda(t_0)\dot x(t_0)+\dot\lambda_0(t_0)\Big)\bar t_0^2 + 2\dot\lambda(t_0)\bar x(t_0)\bar t_0$$
$$-\Big(\dot\lambda(t_f)\dot x(t_f)+\dot\lambda_0(t_f)\Big)\bar t_f^2 - 2\dot\lambda(t_f)\bar x(t_f)\bar t_f. \tag{45}$$

5.4 Second-Order Sufficient Optimality Conditions (SSC)

The state variable x_i is called *unessential* if the function f does not depend on x_i and the functions F,J,K are affine in $x_{i0} := x_i(t_0)$ and $x_{if} := x_i(t_f)$. Let $\underline{x}$ denote the vector of all essential components of state vector x. Let us define a strong minimum in the basic problem.

Definition 12. The process $\mathscr{T}$ affords *a strong minimum* if there exists $\varepsilon > 0$ such that $\mathscr{J}(\tilde{\mathscr{T}}) \ge \mathscr{J}(\mathscr{T})$ for all admissible processes $\tilde{\mathscr{T}} = (\tilde x(t),\tilde u(t) \mid t\in[\tilde t_0,\tilde t_f])$ satisfying the conditions

(*a*) $|\tilde t_0 - t_0| < \varepsilon, \quad |\tilde t_f - t_f| < \varepsilon, \quad |\tilde x(\tilde t_0) - x(t_0)| < \varepsilon,$
(*b*) $\max\limits_{\tilde\Delta\cap\Delta} |\underline{\tilde x}(t) - \underline{x}(t)| < \varepsilon, \quad$ where $\tilde\Delta = [\tilde t_0,\tilde t_f],$

The *strict* strong minimum is defined in a similar way, with the non-strict inequality $\mathscr{J}(\tilde{\mathscr{T}}) \ge \mathscr{J}(\mathscr{T})$ replaced by the strict one and the process $\tilde{\mathscr{T}}$ required to be different from $\mathscr{T}$.

A natural strengthening of the necessary Condition $\mathscr{A}_\Theta$ of Theorem 13 turns out to be a sufficient optimality condition not only for a Pontryagin minimum, but also for a strong minimum.

Theorem 14. *Let the following Condition $\mathscr{B}_\Theta$ be fulfilled for the process $\mathscr{T}$:*

(a) $u(t)$ is a strict bang-bang control (i.e., there exists $\mu\in M_0$ such that condition (35) holds),
(b) there exists $\mu\in M_0$ such that $D^k(H^\mu) > 0$, $k = 1,\dots,s$,
(c) $\max\limits_{\mu\in M_0} \Omega_\Theta(\mu,\bar z) > 0$ *for all* $\bar z \in \mathscr{K}_\Theta \setminus \{0\}$.

Then $\mathscr{T}$ is a strict strong minimum.

Note that the condition (c) is automatically fulfilled if $\mathscr{K}_\Theta = \{0\}$, which gives a *first-order sufficient condition* for a strong minimum in the problem. Also note that the condition (c) is automatically fulfilled if there exists $\mu\in M_0$ such that

$$\Omega_\Theta(\mu,\bar{z}) > 0 \text{ for all } \bar{z} \in \mathscr{K}_\Theta \setminus \{0\}. \tag{46}$$

Sufficient conditions for inequality (46) were obtained in [21] and [22] (see also [36], Section 6.3). Clearly, there is *no gap* between the necessary condition $\mathscr{A}_\Theta$ of Theorem 13 and the sufficient condition $\mathscr{B}_\Theta$ of Theorem 14.

6 Induced Optimization Problem for Bang-Bang Controls and the Verification of SSC

We continue our discussion of bang-bang controls. Second-order sufficient optimality conditions for bang-bang controls had been derived in the literature in two different forms. The first form was discussed in the last section. The second one is due to Agrachev et al. [1], who first reduce the bang-bang control problem to a finite-dimensional optimization problem and then show that the well-known sufficient optimality conditions for this optimization problem supplemented by the strict bang-bang property furnish sufficient conditions for the bang-bang control problem. The bang-bang control problem, considered in this section, is more general than that in [1]. Following [35], we claim the equivalence of both forms of second-order conditions for this problem.

6.1 Formulation of the Induced Optimization Problem and Necessary Optimality Conditions

Let $\widehat{\mathscr{T}} = (\widehat{x}(t),\widehat{u}(t) \mid t \in [\widehat{t}_0,\widehat{t}_f])$ be an admissible process for the basic control problem (22)–(24). We denote by ex U the set of vertices of the polyhedron U. Assume that $\widehat{u}(t)$ is a bang-bang control in $\widehat{\Delta} = [\hat{t}_0,\hat{t}_f]$ taking values in the set ex U,

$$\hat{u}(t) = u^k \in \text{ex } U \quad \text{for } t \in (\hat{t}_{k-1},\hat{t}_k), \quad k = 1,\dots,s+1,$$

where $\hat{t}_{s+1} = \hat{t}_f$. Thus, $\widehat{\Theta} = \{\hat{t}_1,\dots,\hat{t}_s\}$ is the set of switching points of the control $\hat{u}(\cdot)$ with $\hat{t}_k < \hat{t}_{k+1}$ for $k = 0,1,\dots,s$. Assume now that the set M_0 of multipliers is nonempty for the process $\widehat{\mathscr{T}}$. Put

$$\hat{x}(\hat{t}_0) = \hat{x}_0, \qquad \hat{\theta} = (\hat{t}_1,\dots,\hat{t}_s), \qquad \hat{\zeta} = (\hat{t}_0,\hat{t}_f,\hat{x}_0,\hat{\theta}). \tag{47}$$

Then $\hat{\theta} \in \mathbb{R}^s$, $\hat{\zeta} \in \mathbb{R}^2 \times \mathbb{R}^n \times \mathbb{R}^s$, where $n = d(x)$.

Take a small neighborhood $\mathscr{V}$ of the point $\hat{\zeta}$ in $\mathbb{R}^2 \times \mathbb{R}^n \times \mathbb{R}^s$, and let

$$\zeta = (t_0,t_f,x_0,\theta) \in \mathscr{V},$$

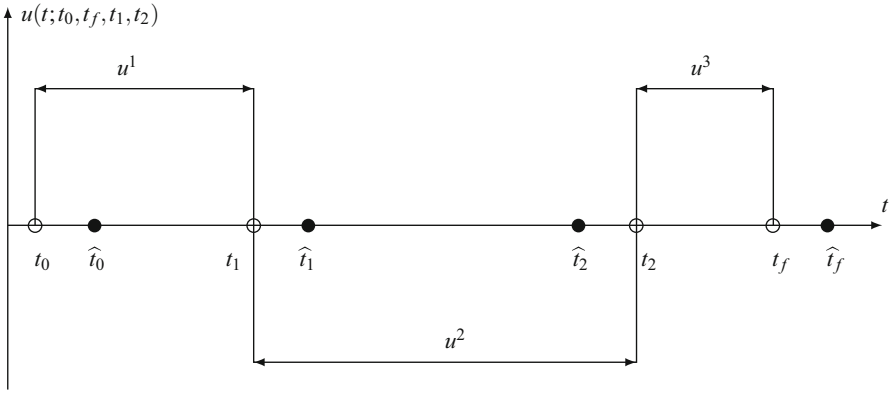

Fig. 2 Bang-bang control with two switches

where $\theta = (t_1, \dots, t_s)$ satisfies $t_0 < t_1 < t_2 < \dots < t_s < t_f$. Define the function $u(t;t_0,t_f,\theta)$ by the condition (Fig. 2)

$$u(t;t_0,t_f,\theta) = u^k \quad \text{for } t \in (t_{k-1},t_k), \quad k = 1,\dots,s+1, \tag{48}$$

where $t_{s+1} = t_f$. The values $u(t_k;t_0,t_f,\theta)$, $k = 1,\dots,s$, may be chosen in U arbitrarily. For definiteness, define them by the condition of continuity of the control from the left: $u(t_k;t_0,t_f,\theta) = u(t_k-;t_0,t_f,\theta)$, $k = 1,\dots,s$.

Let $x(t;t_0,t_f,x_0,\theta)$ be the solution of the initial value problem (IVP)

$$\dot{x} = f(t,x) + g(t,x)u(t;t_0,t_f,\theta), \quad t \in [t_0,t_f], \quad x(t_0) = x_0. \tag{49}$$

For each $\zeta \in \mathcal{V}$ this solution exists if the neighborhood $\mathcal{V}$ of the point $\hat{\zeta}$ is sufficiently small. We obviously have

$$x(t;\hat{t}_0,\hat{t}_f,\hat{x}_0,\hat{\theta}) = \hat{x}(t), \quad t \in \widehat{\Delta}, \quad u(t;\hat{t}_0,\hat{t}_f,\hat{\theta}) = \hat{u}(t), \quad t \in \widehat{\Delta} \setminus \widehat{\Theta}.$$

Consider now the following finite-dimensional optimization problem in the space $\mathbb{R}^2 \times \mathbb{R}^n \times \mathbb{R}^s$ of the variables $\zeta = (t_0,t_f,x_0,\theta)$:

$$\begin{aligned} &\mathscr{J}(\zeta) := J(t_0,x_0,t_f,x(t_f;t_0,t_f,x_0,\theta)) \to \min, \\ &\mathscr{F}(\zeta) := F(t_0,x_0,t_f,x(t_f;t_0,t_f,x_0,\theta)) \le 0, \\ &\mathscr{G}(\zeta) := K(t_0,x_0,t_f,x(t_f;t_0,t_f,x_0,\theta)) = 0. \end{aligned} \tag{50}$$

We call (50) the *Induced Optimization Problem* (IOP) or simply *Induced Problem* which represents an extension of the IOP introduced in [1]. The following assertion is almost obvious.

Theorem 15. *Let the process $\widehat{\mathscr{T}}$ be a Pontryagin local minimum for the basic control problem (22)–(24). Then the point $\hat{\zeta}$ is a local minimum of the IOP (50), and hence it satisfies first and second-order necessary conditions for this problem.*

6.2 Second-Order Optimality Conditions for Bang-Bang Controls in Terms of the Induced Optimization Problem

We shall clarify a relationship between the second-order conditions for the Induced Optimization Problem (50) at the point $\hat{\zeta}$ and those in the basic bang-bang control problem (22)–(24) for the process $\widehat{\mathscr{T}}$. It turns out that there is a one-to-one correspondence between Lagrange multipliers in these problems and a one-to-one correspondence between elements of the critical cones. Moreover, for corresponding Lagrange multipliers, the quadratic forms in these problems take equal values on the corresponding elements of the critical cones. This allows to express the necessary and sufficient quadratic optimality conditions for a bang-bang control, formulated in Theorems 13 and 14, in terms of the IOP (50). Thus we are able to establish the equivalence between our quadratic sufficient conditions and those due to Agrachev et al. [1].

Let $\widehat{\mathscr{T}} = (\hat{x}(t), \hat{u}(t) \mid t \in [\hat{t}_0, \hat{t}_f])$ be an admissible process in the basic problem with the properties assumed in Sect. 5.2 and let $\hat{\zeta} = (\hat{t}_0, \hat{t}_f, \hat{x}_0, \hat{\theta})$ be the corresponding admissible point in the IOP. The Lagrange function for the Induced Optimization Problem (50) is

$$\mathscr{L}(\mu, \zeta) = \mathscr{L}(\mu, t_0, t_f, x_0, \theta) = \alpha_0 \mathscr{J}(\zeta) + \alpha \mathscr{F}(\zeta) + \beta \mathscr{G}(\zeta), \tag{51}$$

where $\mu = (\alpha_0, \alpha, \beta)$, $\zeta = (t_0, t_f, x_0, \theta)$, $\theta = (t_1, \ldots, t_s)$. We denote by $\mathscr{K}_0$ the critical cone at the point $\hat{\zeta}$ in the IOP. Thus, $\mathscr{K}_0$ is the set of collections $\bar{\zeta} = (\bar{t}_0, \bar{t}_f, \bar{x}_0, \bar{\theta})$ such that

$$\mathscr{J}'(\hat{\zeta})\bar{\zeta} \le 0, \quad \mathscr{F}_i'(\hat{\zeta})\bar{\zeta} \le 0,\ i \in I, \quad \mathscr{G}'(\hat{\zeta})\bar{\zeta} = 0, \tag{52}$$

where $I = \{i \mid \mathscr{F}_i(\hat{\zeta}) = 0\}$ is the set of indices of the inequality constraints active at the point $\hat{\zeta}$. For $\mu \in M_0$ the quadratic form, of the induced optimization problem, is equal to $\langle \mathscr{L}_{\zeta\zeta}(\mu, \hat{\zeta})\bar{\zeta}, \bar{\zeta} \rangle$.

Let us formulate now second-order optimality conditions for the basic control problem in terms of the IOP.

Theorem 16 (Second-Order Necessary Conditions). *If the process $\widehat{\mathscr{T}}$ affords a Pontryagin minimum in the basic problem, then the following Condition $\mathscr{A}_0$ holds: the set M_0 is nonempty and*

$$\max_{\mu \in M_0} \langle \mathscr{L}_{\zeta\zeta}(\mu, \hat{\zeta})\bar{\zeta}, \bar{\zeta} \rangle \ge 0 \quad \textit{for all} \quad \bar{\zeta} \in \mathscr{K}_0.$$

Theorem 17 (Second-Order Sufficient Conditions). *Let the following Condition $\mathscr{B}_0$ be fulfilled for an admissible process $\widehat{\mathscr{T}}$ in the basic problem:*

(a) $\hat{u}(t)$ is a strict bang-bang control with finitely many switching times $\hat{t}_k$, $k = 1,\dots,s$ (hence, the set M_0 is nonempty and condition (35) holds for some $\mu \in M_0$),
(b) there exists $\mu \in M_0$ such that $D^k(H^\mu) > 0$, $k = 1,\dots,s$,
(c) $\max\limits_{\mu\in M_0} \langle \mathscr{L}_{\zeta\zeta}(\mu,\hat{\zeta})\bar{\zeta},\bar{\zeta}\rangle > 0$ for all $\bar{\zeta} \in \mathscr{K}_0 \setminus \{0\}$.

Then $\widehat{\mathscr{T}}$ is a strict strong minimum in the basic problem.

Theorem 17 is a generalization of sufficient optimality conditions for bang-bang controls obtained in Agrachev et al. [1]. Detailed proofs of Theorems 16 and 17 are given in [35] and in our book [36].

Remark 2. Noble and Schättler [29] and Schättler and Ledzewicz [38] develop sufficient optimality conditions using methods of geometric optimal control. There is some evidence that their sufficient conditions are closely related to the SSC in our work [35, 36]. However, a formal proof of the equivalence of both types of sufficient conditions has not yet been worked out.

6.3 Numerical Methods for Solving the Induced Optimization Problem

The arc-parametrization method developed in [16, 26] provides an efficient method for solving the IOP. To better explain this method, for simplicity let us consider the basic control problem with fixed initial time $t_0 = 0$ and fixed initial condition $x_0(0) = x_0$, and without inequality constraints $\mathscr{F}(\zeta) \le 0$. For this problem, we slightly change the notation and replace the resulting optimization vector $\zeta = (t_f, t_1, \dots, t_s)$ by the vector $z = (t_1,\dots,t_s,t_{s+1}), t_{s+1} = t_f$. Instead of directly optimizing the switching times t_k , $k = 1,\dots,s$, we determine the *arc lengths* (arc durations)

$$\xi_k := t_k - t_{k-1}, \quad k = 1,\dots,s,s+1, \tag{53}$$

of bang-bang arcs. Hence, the optimization variable $z = (t_1,\dots,t_s,t_{s+1})^*$ is replaced by the optimization variable

$$\xi := (\xi_1,\dots,\xi_s,\xi_{s+1})^* \in \mathbb{R}^{s+1}, \quad \xi_k := t_k - t_{k-1}. \tag{54}$$

The variables z and ξ are related by a linear transformation involving the regular $(s+1)\times(s+1)$-matrix R :

$$\xi = Rz, \quad z = R^{-1}\xi, \quad R = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{pmatrix}. \tag{55}$$

In the arc-parametrization method, the time interval $[t_{k-1}, t_k]$ is mapped to the fixed interval

$$I_k := \left[\frac{k-1}{s+1}, \frac{k}{s+1}\right], \quad k = 1, \dots, s+1, \tag{56}$$

by the linear transformation

$$t = a_k + b_k\tau, \quad \tau \in I_k, \tag{57}$$

where

$$a_k = t_{k-1} - (k-1)\xi_k, \quad b_k = (s+1)\xi_k. \tag{58}$$

Identifying

$$x(\tau) \cong x(a_k + b_k\tau) = x(t) \tag{59}$$

in the relevant intervals, we obtain the ODE system

$$\dot{x}(\tau) = (s+1)\xi_k(f(a_k + b_k\tau, x(\tau)) + g(a_k + b_k\tau, x(\tau))u^k) \quad \text{for} \quad \tau \in I_k. \tag{60}$$

By concatenating the solutions in the intervals I_k we get the continuous solution $x(t) = x(t;\xi)$ in the normalized interval $[0,1]$. When expressed via the new optimization variable ξ, the Induced Optimization Problem (IOP) in (50) is equivalent to the following optimization problem ($\widetilde{\text{IOP}}$) with $t_f = \sum_{k=1}^{s+1} \xi_k$:

$$\begin{aligned} &\text{Minimize} \quad \widetilde{\mathscr{J}}(\xi) := J(t_f; x(1,\xi)) \\ &\text{subject to} \quad \widetilde{\mathscr{G}}(\xi) := K(t_f; x(1,\xi)) = 0. \end{aligned} \tag{61}$$

The Lagrangian function is given by

$$\widetilde{\mathscr{L}}(\mu, \xi) = \alpha_0 \widetilde{\mathscr{J}}(\xi) + \beta\widetilde{\mathscr{G}}(\xi), \quad \mu = (\alpha_0, \beta). \tag{62}$$

Using the linear transformation (55), it can easily be seen that the SSCs for the Induced Optimization Problems (IOP) and ($\widetilde{\text{IOP}}$) are equivalent; cf. similar

arguments in [26]. To solve the $(\widetilde{\mathrm{IOP}})$, we use a suitable adaptation of the control package NUDOCCCS in Büskens [4, 5]. Then we can take advantage of the fact that NUDOCCCS also provides the Jacobian of the terminal constraints and the Hessian of the Lagrangian which are needed in the check of the second-order condition in Theorem 17.

In practice, we shall verify the positive definiteness condition (c) in Theorem 17 in a stronger form. We assume that the multiplier can be chosen as $\mu = (1, \beta)$ and that the following *regularity condition* holds:

$$\operatorname{rank} \widetilde{\mathscr{G}}_\xi(\hat{\xi}) = d(K).$$

Let N be the $n_\xi \times (n_\xi - d(K))$ matrix, $n_\xi = n + s + 1$ (where $n = d(x)$), with full column rank $n_\xi - d(K)$, whose columns span the kernel of $\widetilde{\mathscr{G}}_\xi(\hat{\xi})$. Then condition (c) in Theorem 17,

$$\langle \widetilde{\mathscr{L}}_{\xi\xi}(\hat{\xi}, \beta)\bar{\xi}, \bar{\xi}\rangle > 0 \quad \forall \bar{\xi} \neq 0, \quad \widetilde{\mathscr{G}}_\xi(\hat{\xi})\bar{\xi} = 0, \tag{63}$$

is equivalent to the condition that the *projected Hessian* is positive definite [6],

$$N^* \widetilde{\mathscr{L}}_{\xi\xi}(\hat{\xi}, \beta) N > 0\,. \tag{64}$$

7 Numerical Example with Fixed Final Time: Optimal Control of the Chemotherapy of HIV

The treatment of patients infected with the human immunodeficiency virus (HIV) is still of great concern today (Kirschner et al. [18]). The problem of determining optimal chemotherapies has been treated in Kirschner et al. [18] in the framework of optimal control theory. The optimal control model is based on a simple dynamic model in Perelson et al. [37] which simulates the interaction of the immune system with HIV. Kirschner et al. [18] use a control quadratic cost functional of L^2-type. It has been argued in Schättler et al. [39] that in a biological context it is more appropriate to consider cost functionals of L^1-type which are linear in the control variable. Therefore, in this section, we are studying an objective of L^1-type, where the quadratic control is replaced by a linear control. The state and control variables have the following meaning:

$T(t)$: concentration of uninfected $CD4^+$ T cells,
$T^*(t)$: concentration of latently infected $CD4^+$ T cells,
$T^{**}(t)$: concentration of actively infected $CD4^+$ T cells,
$V(t)$: concentration of free infectious virus particles,
$u(t)$: control, rate of chemotherapy.

The treatment starts at $t_0 = 0$ and terminates at the fixed final time $t_f = 500$ (days). Thus, the control process is considered in the interval $[0, t_f]$. The dynamics of the populations are (omitting the time argument):

$$\begin{aligned} dT/dt &= \tfrac{s}{1+V} - \mu_T T + rT\left(1 - \tfrac{T+T^*+T^{**}}{T_{\max}}\right) - k_1 VT\,, \quad T(0) = T_0, \\ dT^*/dt &= k_1 VT - \mu_T T^* - k_2 T^*\,, \qquad\qquad\qquad\quad T^*(0) = T_0^*, \\ dT^{**}/dt &= k_2 T^* - \mu_b T^{**}, \qquad\qquad\qquad\qquad T^{**}(0) = T_0^{**} \\ dV/dt &= (1-u) N \mu_b T^{**} - k_1 VT - \mu_V V\,, \qquad\quad V(0) = V_0\,. \end{aligned} \tag{65}$$

The control constraint is given by

$$0 \le u(t) \le 1 \quad \forall\, t \in [0, t_f], \tag{66}$$

where $u(t) = 1$ represents *maximal* chemotherapy, while $u(t) = 0$ means that no chemotherapy is administered. Note that Kirschner et al. [18] consider the control variable $v = 1 - u$. It is convenient to write the ODE (65) as the control affine system (24),

$$\dot{x} = f(x) + g(x)u, \quad x(0) = x_0, \quad x = (T, T^*, T^{**}, V) \in \mathbb{R}^4, \tag{67}$$

with obvious definitions of the vector functions $f(x)$ and $g(x)$. As in [18] we consider two sets of initial conditions which depend on the time at which the treatment starts after the infection. The following initial conditions are interpolated from [18] and have already been used in the [15].

Initial conditions after 800 days:

$$T_0 = 982.8, \quad T_0^* = 0.05155, \quad T_0^{**} = 0.0006175\,, \quad V_0 = 0.07306. \tag{68}$$

Initial conditions after 1000 days:

$$T_0 = 904.1, \quad T_0^* = 0.3447, \quad T_0^{**} = 0.004167, \quad V_0 = 0.4939. \tag{69}$$

The parameter and constants are taken from [18] and are listed in Table 1.

Kirschner et al. [18] consider the following objective of L^2-type which is quadratic in the control variable:

$$\text{Minimize} \quad J(x,u) = \int_0^{t_f} (-T(t) + B\,u(t)^2)\,dt \quad (B = 50). \tag{70}$$

Recall that the state variable is defined as $x := (T, T^*, T^{**}, V) \in \mathbb{R}^4$. The optimal control that minimizes (70) subject to the constraints (65)–(69) is a *continuous*

Table 1 Parameters and constants

		Parameters and constants	Values
μ_T	:	death rate of uninfected CD4$^+$ T cell population	0.02 d^{-1}
μ_{T^*}	:	death rate of latently infected CD4$^+$ T cell population	0.02 d^{-1}
μ_b	:	death rate of actively infected CD4$^+$ T cell population	0.24 d^{-1}
μ_V	:	death rate of free virus	2.4 d^{-1}
k_1	:	rate CD4$^+$ T cells becomes infected by free virus	2.4×10^{-5} mm^3 d^{-1}
k_2	:	rate T* cells convert to actively infected	3×10^{-3} mm^3 d^{-1}
r	:	rate of growth for the CD4$^+$ T cell population	0.03 d^{-1}
N	:	number of free virus produced by T^{**} cells	1200
$T_{\max}$	:	maximum CD4$^+$ T cell population level	1.5×10^3 mm^{-3}
s	:	source term for uninfected CD4$^+$ T cells,	10 d^{-1} mm^{-3}
		where s is the parameter in the source term	$s/(1+V)$

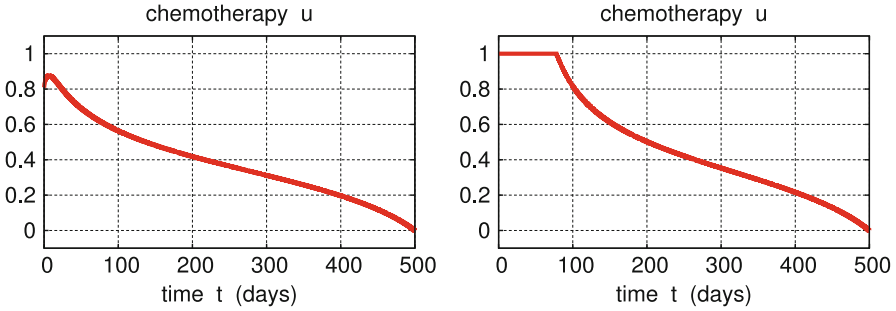

Fig. 3 Optimal control for the L^2 functional (70). (*Left*) begin of treatment after 800 days: initial conditions (68). (*Right*) begin of treatment after 1000 days: initial conditions (69)

function, since the associated Hamiltonian $H(x,\lambda,u)$ has a unique minimum with respect to u and the strict Legendre-Clebsch condition $H_{uu} = 2B > 0$ holds. For the two sets of initial conditions (68) and (69), the numerical discretization and NLP approach using AMPL [12] and IPOPT [42] yield the optimal controls shown in Fig. 3 which were also obtained in Hannemann [15].

Hannemann [15] showed that second-order sufficient conditions (SSC) are satisfied for the controls displayed in Fig. 3, since the associated matrix Riccati equation has a bounded solution. Note that Riccati equations are discussed in [19, 24] and in our book [36], Chap. 4.

Instead of the L^2 functional (70) we consider now the functional of L^1-type:

$$\text{Minimize} \quad J_1(x,u) = \int_0^{t_f} (-T(t) + Bu(t))\,dt \quad (B = 50). \tag{71}$$

The Hamiltonian or Pontryagin function for this control problem is given by

$$H(x,\lambda,u) = -T + Bu + \lambda\,(f(x) + g(x)u)\,, \tag{72}$$

where $\lambda = (\lambda_T, \lambda_{T^*}, \lambda_{T^{**}}, \lambda_V) \in \mathbb{R}^4$ denotes the adjoint variable. The adjoint equation and transversality condition are given by

$$\dot{\lambda} = -H_x(x, \lambda, u), \quad \lambda(t_f) = (0,0,0,0),$$

since the terminal state $x(t_f)$ is free and the objective (71) does not contain a Mayer term. We do not write out the adjoint equation $\dot{\lambda} = -H_x(x, \lambda, u)$ explicitly, since this equation is not needed in the sequel. The adjoint variables can be computed from the Lagrange multipliers of the associated Induced Optimization Problem (IOP). The *switching function* is given by

$$\sigma(x, \lambda) = H_u(x, \lambda, u) = B - \lambda_V N \mu_b T^{**}, \quad \sigma(t) = \sigma(x(t), \lambda(t)). \tag{73}$$

The minimization of the Hamiltonian with respect to u yields the switching condition

$$u(t) = \left\{ \begin{array}{l} 1, \text{ if } \sigma(t) < 0 \\ 0, \text{ if } \sigma(t) > 0 \end{array} \right\}. \tag{74}$$

The control has a *singular arc* in an interval $[t_1, t_2] \subset [0, T]$ if $\sigma(t) = 0$ holds on $[t_1, t_2]$. However, we do not discuss singular controls further because for the data in Table 1 we never found singular arcs. Indeed, the optimal control for the L^1-functional (71) is the following bang-bang control with only one switch at t_1 :

$$u(t) = \left\{ \begin{array}{l} 1 \text{ for } 0 \le t < t_1 \\ 0 \text{ for } t_1 \le t \le t_f \end{array} \right\} \tag{75}$$

The terminal arc $u(t) = 0$ results from the terminal value $\sigma(t_f) = B > 0$ of the switching function and the minimum condition (74). Hence, the IOP has only the scalar optimization variable t_1 and thus the objective (71) reduces to a function $J_1(t_1) = J_1(x, u)$. The arc-parametrization method [26, 36] and the code NUDOCCCS [4] yield the following numerical results, where state variables are listed with 8 digits and adjoint variables with 8 digits.

$$\begin{array}{lll} J_1 = -489810.5, & t_1 = 161.6957, & T(t_f) = 983.4926, \\ T^*(t_f) = 0.04934668, & T^{**}(t_f) = 0.0005910497, & V(t_f) = 0.06993300, \\ \lambda_T(0) = -0.125173, & \lambda_{T^*}(0) = -1.51988, & \lambda_{T^{**}}(0) = -2.94704, \\ \lambda_V(0) = -0.449700. & & \end{array} \tag{76}$$

The state and control variables and the switching function are displayed in Fig. 4. To verify that the second-order sufficient conditions (SSC) are satisfied for the computed extremal solution, we have to check the conditions of Theorem 17. The strict bang-bang property is satisfied, since we infer from Fig. 4 (bottom, right) that the switching function satisfies

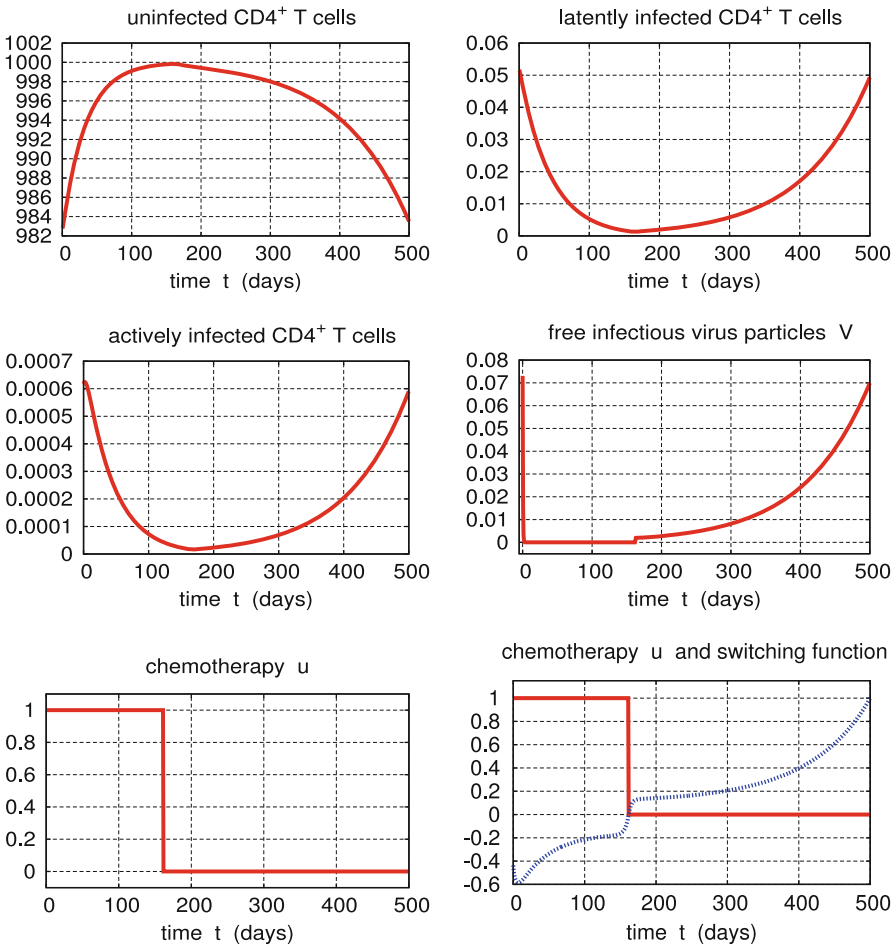

Fig. 4 Optimal solution for initial conditions (68): treatment starts after 800 days. *Top row:* (*left*) uninfected $CD4^+T$ cells, (*right*) latently infected $CD4^+T^*$ cells. *Middle row:* (*left*) actively infected $CD4^+T^{**}$ cells, (*right*) infectious virus particles V. *Bottom row:* (*left*) bang-bang control u, (*right*) bang-bang control u and (scaled) switching function σ in (73) satisfying the switching condition (74)

$$\sigma(t) < 0 \quad \text{for} \quad 0 \le t < t_1\,, \quad \dot\sigma(t_1) > 0, \quad \sigma(t) > 0 \quad \text{for} \quad t_1 < t \le t_f = 500. \tag{77}$$

To verify the positive definiteness in condition (63), we note that the Hessian is simply the second derivative of the objective $J_1(t_1)$ evaluated at the optimal switching time $t_1 = 161.695711$ for which we find

$$\frac{d^2 J_1}{dt_1^2} = 1.5469 > 0.$$

Hence, the extremal solution (76) displayed in Fig. 4 provides a strict strong minimum.

Now we try to improve the optimal terminal value $T(t_f) = 983.493$ of the uninfected $CDC4^+T$ cells. For that purpose we prescribe a higher terminal value and minimize the functional $J_1(x,u)$ subject to the boundary condition

$$T(t_f) = 995\,. \tag{78}$$

The arc-parametrization method [26, 36] and the control package NUDOCCCS [4] furnish the results

$$\begin{aligned} J_1 &= -489044.529, & t_1 &= 198.566451, & T(t_f) &= 995.0, \\ T^*(t_f) &= 0.014576433, & T^{**}(t_f) &= 0.00017436211, & V(t_f) &= 0.020625442, \\ \lambda_T(0) &= -33.7027, & \lambda_T(t_f) &= -211.377, & \lambda_{T^*}(0) &= 28078.4 \\ \lambda_{T^{**}}(0) &= 2.76312, & \lambda_V(0) &= 405.843. \end{aligned} \tag{79}$$

Figure 5 displays the state and control variables and the switching function. Figure 5 (bottom, right) shows that the strict bang-bang property (77) is satisfied. Condition *(c)* in Theorem 17 holds because the critical cone $\mathscr{K}_0 = \{0\}$ contains of zero element. Therefore, the extremal displayed in Fig. 5 provides a strict strong minimum.

Finally, we study the optimal solution for the initial values (69), when the treatment starts after 1000 days and, again, the boundary condition $T(t_f) = 995$ is prescribed. The arc-parametrization method in [26, 36] and the control package NUDOCCCS yield the results

$$\begin{aligned} J_1 &= -483480.9, & t_1 &= 254.5443, & T(t_f) &= 995.0, \\ T^*(t_f) &= 0.01457694, & T^{**}(t_f) &= 0.0001743682, & V(t_f) &= 0.02062616, \\ \lambda_T(0) &= -35.629, & \lambda_T(t_f) &= -214.021, & \lambda_{T^*}(0) &= 4209.98, \\ \lambda_{T^{**}}(0) &= 2.69187, & \lambda_V(0) &= 136.835. \end{aligned} \tag{80}$$

Figure 6 depicts the state and control variables and the switching function. The SSC in Theorem 17 are satisfied, since the *strict bang-bang property* (77) holds and condition (63) holds in view of $\mathscr{K}_0 = \{0\}$. Therefore, the extremal (80) provides a strict strong minimum.

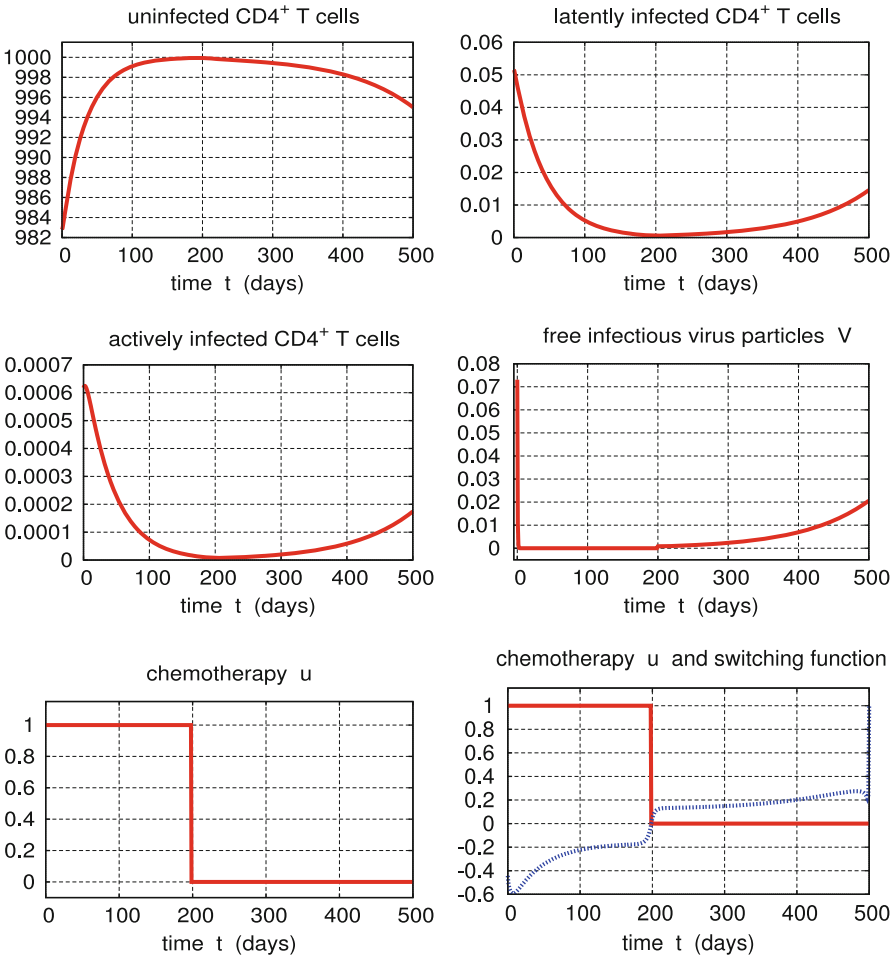

Fig. 5 Optimal solution for initial conditions (68): treatment starts after 800 days and and terminal condition $T(t_f) = 995$. *Top row:* (*left*) uninfected $CD4^+T$ cells, (*right*) latently infected $CD4^+T^*$. *Middle row:* (*left*) actively infected $CD4^+T^{**}$ cells, (*right*) infectious virus particles V. *Bottom row:* (*left*) bang-bang control u, (*right*) bang-bang control u and scaled switching function σ in (73) satisfying the switching condition (74)

8 Numerical Example with Free Final Time: Time-Optimal Control of Two-Link Robots

In this section, we review the results in our book [36] on the optimal control of two-link robots which has been addressed in various articles; cf., e.g. [9, 13, 14, 30]. In these papers, optimal control policies are determined solely on the basis of first

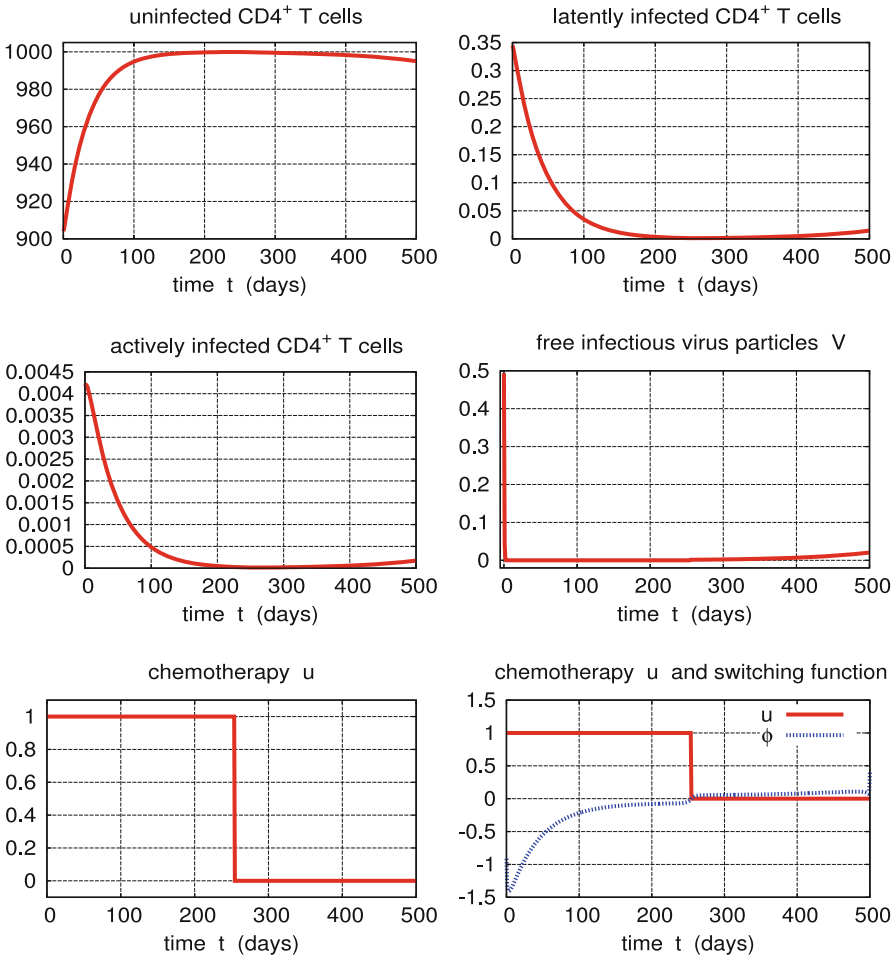

Fig. 6 Optimal solution for initial conditions (69): treatment starts after 1000 days and terminal condition $T(t_f) = 995$. *Top row:* (*left*) uninfected $CD4^+T$ cells, (*right*) latently infected $CD4^+T^*$. *Middle row:* (*left*) actively infected $CD4^+T^{**}$ cells, (*right*) infectious virus particles V. *Bottom row:* (*left*) bang-bang control u, (*right*) bang-bang control u and scaled switching function σ in (73) satisfying the switching condition (74)

order necessary conditions, since sufficient conditions were not available. In this section we show that SSC hold for both types of robots considered in [9, 14, 30].

First, we study the robot model considered in Chernousko et al. [9]. Göllmann [14] has shown that the optimal control candidate presented in [9] is not optimal, since the sign conditions of the switching functions do not comply with the Minimum Principle. Figure 7 displays the two-link robot schematically. The state variables are the angles q_1 and q_2. The parameters I_1 and I_2 are the moments of inertia of the upper arm $\overline{OQ}$ and the lower arm $\overline{QP}$ with respect to the points O and

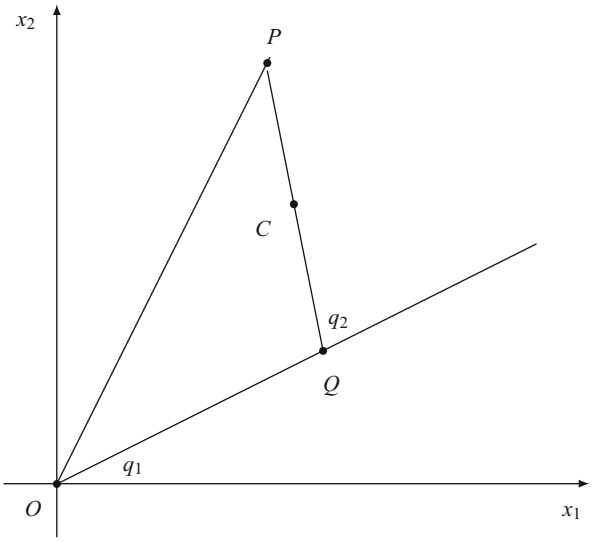

Fig. 7 Schematical representation of a two-link robot

Q, resp. Further, let m_2 be the mass of the lower arm, $L_1 = |\overline{OQ}|$ the length of the upper arm, and $L_1 = |\overline{QC}|$ the distance between the second link Q and the center of gravity C of the lower arm. With the abbreviations

$$\begin{aligned} A &= I_1 + m_2 L_1^2 + I_2 + 2m_2 L_1 L \cos q_2\,, & B &= I_2 + m_2 L_1 L \cos q_2\,, \\ R_1 &= u_1 + m_2 L_1 L(2\dot q_1 + \dot q_2)\dot q_2 \sin q_2\,, & R_2 &= u_2 - m_2 L_1 L \dot q_1^2 \sin q_2\,, \\ D &= I_2\,, & \Delta &= AD - B^2\,, \end{aligned} \tag{81}$$

the dynamics of the two-link robot can be described by the ODE system

$$\begin{aligned} \dot q_1 &= \omega_1\,, \quad & \dot\omega_1 &= \tfrac{1}{\Delta}(DR_1 - BR_2)\,, \\ \dot q_2 &= \omega_2\,, \quad & \dot\omega_2 &= \tfrac{1}{\Delta}(AR_2 - BR_1)\,, \end{aligned} \tag{82}$$

where ω_1 and ω_2 are the angular velocities. The torques u_1 and u_2 in the two links represent the two control variables. Therefore, the state variable and control variable are given by

$$x = (q_1, q_2, \omega_1, \omega_2) \in \mathbb{R}^4, \quad u = (u_1, u_2) \in \mathbb{R}^2.$$

The control problem consists in steering the robot from a given initial position to a terminal position in minimal final time t_f,

$$\begin{aligned} &q_1(0) = 0\,, \quad && q_2(0) = 0\,, \quad && \omega_1(0) = 0\,, \ \omega_2(0) = 0\,, \\ &q_1(t_f) = -0.44\,, \ && q_2(t_f) = 1.83\,, \ && \omega_1(t_f) = 0\,, \ \omega_2(t_f) = 0\,. \end{aligned} \tag{83}$$

The control components are bounded by

$$|u_1(t)| \le 2\,, \qquad |u_2(t)| \le 1\,, \qquad t \in [0, t_f]\,. \tag{84}$$

The Pontryagin function (Hamiltonian) is

$$H = \lambda_1 \omega_1 + \lambda_2 \omega_2 + \frac{\lambda_3}{\Delta}\left(DR_1(u_1) - BR_2(u_2)\right) + \frac{\lambda_4}{\Delta}\left(AR_2(u_2) - BR_1(u_1)\right)\,. \tag{85}$$

The adjoint equations are rather complicated and are not given here explicitly. The switching functions are

$$\sigma_1(x,\lambda) = H_{u_1} = \frac{\lambda_3}{\Delta} D - \frac{\lambda_4}{\Delta} B\,, \quad \sigma_2(x,\lambda) = H_{u_2} = \frac{\lambda_4}{\Delta} A - \frac{\lambda_3}{\Delta} B\,. \tag{86}$$

For the parameter values

$$L_1 = 1\,, \quad L = 0.5\,, \quad m_2 = 10\,, \quad I_1 = I_2 = \frac{10}{3}\,,$$

Göllmann [14] has found the following control structure with four bang-bang arcs,

$$u(t) = (u_1(t), u_2(t)) = \left\{ \begin{array}{r} (-2,1)\ ,\ 0 \le t < t_1 \\ (2,1)\ ,\ t_1 \le t < t_2 \\ (2,-1)\ ,\ t_2 \le t < t_3 \\ (-2,-1)\ ,\ t_3 \le t \le t_f \end{array} \right\},\ 0 < t_1 < t_2 < t_3 < t_f. \tag{87}$$

This control structure differs substantially from the one in Chernousko et al. [9] which violates the switching conditions. Obviously, the bang-bang control (87) satisfies the assumption that only one control components switches at a time. Since the initial point $(q_1(0), q_2(0), \omega_1(0), \omega_2(0))$ is specified, the optimization variable in the IOP (61) is

$$\xi = (\xi_1, \xi_2, \xi_3, \xi_4), \quad \xi_1 = t_1, \quad \xi_2 = t_2 - t_1, \quad \xi_3 = t_3 - t_2, \quad \xi_4 = t_f - t_3\,.$$

Using the code NUDOCCCS we compute the following arc durations and switching times

$$\begin{array}{l} t_1 = 0.7677893\,,\ \xi_2 = 0.3358820\,,\ t_2 = 1.1036713\,, \\ \xi_3 = 1.2626739\,,\ t_3 = 2.3663452\,,\ \xi_4 = 0.8307667\,, \\ t_f = 3.1971119\,. \end{array} \tag{88}$$

Numerical values for the adjoint functions are also provided by the code NUDOCCCS, e.g., the initial values

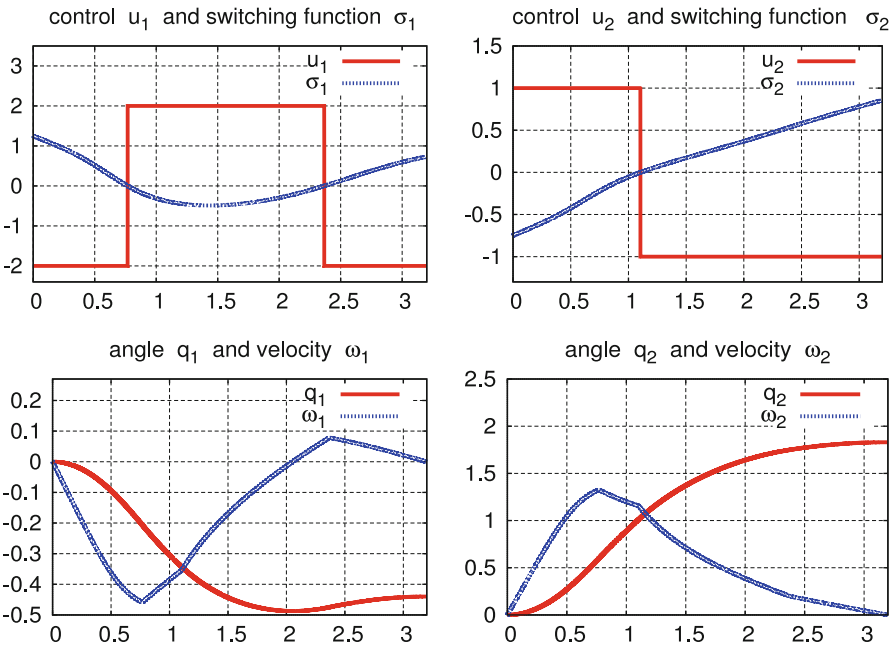

Fig. 8 Control of the two-link robot (81)–(84). *Top row*: (*left*) control u_1 and scaled switching function σ_1, (*right*) control u_2 and scaled switching function σ_2. *Bottom row*: (*left*) angle q_1 and velocity ω_1, (*right*) angle q_2 and velocity ω_2

$$\begin{aligned} \lambda_1(0) &= -1.56972\,,\ \lambda_2(0) = -0.917955\,, \\ \lambda_3(0) &= -2.90537\,,\ \lambda_4(0) = \ -1.45440\,. \end{aligned} \tag{89}$$

Figure 8 shows that the switching functions σ_1 and σ_2 comply with the minimum condition and that the strict bang-bang property (35) and the inequalities $D^k(H) > 0$, $k = 1,2,3$ are satisfied:

$$\begin{aligned} &\sigma_1(t) \neq 0 \quad \text{for} \quad t \neq t_1\,, t_3\,,\ \sigma_2(t) \neq 0 \quad \text{for} \quad t \neq t_2\,, \\ &\dot\sigma_1(t_1) < 0\,, \quad \dot\sigma_1(t_3) > 0\,, \quad \dot\sigma_2(t_2) > 0\,. \end{aligned}$$

For the terminal conditions (83) we obtain the Jacobian

$$\mathscr{G}_\xi(\hat\xi) = \begin{pmatrix} -0.75104 & 0.035106 & 0.25890 & 0 \\ 3.7612 & 1.8493 & -0.20417 & 0 \\ -0.32635 & 0.077005 & 0.21272 & -0.10782 \\ 1.2685 & 0.44545 & -0.48745 & -0.23363 \end{pmatrix}.$$

This square matrix has full rank in view of

$$\det \mathscr{G}_{\xi}(\hat{\xi}) = 0.076652 \neq 0\,,$$

which means that the positive definiteness condition (63) trivially holds. Thus we have verified *first-order* sufficient conditions showing that the extremal solution given by (87)–(89) provides a strict strong minimum.

In the model treated above, some parameters like the mass of the upper arm and the mass of a load at the end of the lower arm appear implicitly in the system equations. The mass m_1 of the upper arm is included in the moment of inertia I_2 and the mass M of a load in the point P can be added to the mass m_2, where the point C and therefore the length L have to be adjusted. The length L_2 of the lower arm is incorporated in the parameter L.

The *second robot model* that we are going to discuss is taken from Geering et al. [13] and Oberle [30]. Here, every physical parameter enters the system equation explicitly. The dynamic system is as follows:

$$\begin{aligned} \dot{q}_1 &= \omega_1\,, & \dot{\omega}_1 &= \tfrac{1}{\Delta}(AI_{22} - BI_{12}\cos q_2)\,, \\ \dot{q}_2 &= \omega_2 - \omega_1\,, & \dot{\omega}_2 &= \tfrac{1}{\Delta}(BI_{11} - AI_{12}\cos q_2)\,, \end{aligned} \tag{90}$$

where we have used the abbreviations

$$\begin{aligned} A &= I_{12}\omega_2^2 \sin q_2 + u_1 - u_2\,, & B &= -I_{12}\omega_1^2 \sin q_2 + u_2\,, \\ \Delta &= I_{11}I_{22} - I_{12}^2 \cos^2 q_2\,, & I_{11} &= I_1 + (m_2 + M)L_1^2\,, \\ I_{12} &= m_2 L L_1 + M L_1 L_2\,, & I_{22} &= I_2 + I_3 + M L_2^2\,. \end{aligned} \tag{91}$$

Here, I_3 denotes the moment of inertia of the load with respect to the point P and ω_2 is now the angular velocity of the angle $q_1 + q_2$. For simplicity, we set $I_3 = 0$. Again, the torques u_1 and u_2 in the two links are used as control variables by which the robot is steered from a given initial position to a non-fixed end position in minimal final time t_f,

$$\begin{aligned} q_1(0) &= 0\,, & \sqrt{(x_1(t_f) - x_1(0))^2 + (x_2(t_f) - x_2(0))^2} &= r\,, \\ q_2(0) &= 0\,, & q_2(t_f) &= 0\,, \\ \omega_1(0) &= 0\,, & \omega_1(t_f) &= 0\,, \\ \omega_2(0) &= 0\,, & \omega_2(t_f) &= 0\,, \end{aligned} \tag{92}$$

where $(x_1(t), x_2(t))$ are the Cartesian coordinates of the point P,

$$\begin{aligned} x_1(t) &= L_1 \cos q_1(t) + L_2 \cos(q_1(t) + q_2(t))\,, \\ x_2(t) &= L_1 \sin q_1(t) + L_2 \sin(q_1(t) + q_2(t))\,. \end{aligned} \tag{93}$$

The initial point $(x_1(0), x_2(0)) = (2, 0)$ is fixed. Both control components are bounded,

$$|u_1(t)| \le 1\,, \qquad |u_2(t)| \le 1\,, \qquad t \in [0, t_f]\,. \tag{94}$$

The Hamilton–Pontryagin function is given by

$$\begin{aligned} H = {} & \lambda_1 \omega_1 + \lambda_2(\omega_2 - \omega_1) + \tfrac{\lambda_3}{\Delta}\,(A(u_1,u_2)I_{22} - B(u_2)I_{12}\cos q_2) \\ & + \tfrac{\lambda_4}{\Delta}\,(B(u_2)I_{11} - A(u_1,u_2)I_{12}\cos q_2)\,. \end{aligned} \tag{95}$$

The switching functions are computed as

$$\begin{aligned} \sigma_1(x,\lambda) &= H_{u_1} = \tfrac{1}{\Delta}\,(\lambda_3 I_{22} - \lambda_4 I_{12}\cos q_2)\,, \\ \sigma_2(x,\lambda) &= H_{u_2} = \tfrac{1}{\Delta}\,(\lambda_3(-I_{22} - I_{12}\cos q_2) + \lambda_4(I_{11} + I_{12}\cos q_2))\,. \end{aligned} \tag{96}$$

For the parameter values

$$L_1 = L_2 = 1\,, \quad L = 0.5\,, \quad m_1 = m_2 = M = 1\,, \quad I_1 = I_2 = \frac{1}{3}\,, \quad I_3 = 0\,, \quad r = 3,$$

we will show that the optimal control has the following structure with five bang-bang arcs with $0 = t_0 < t_1 < t_2 < t_3 < t_4 < t_5 = t_f$ (Fig. 9):

$$u(t) = (u_1(t), u_2(t)) = \left\{ \begin{array}{r} (-1,1) \text{ for } 0 \le t < t_1 \\ (-1,-1) \text{ for } t_1 \le t < t_2 \\ (1,-1) \text{ for } t_2 \le t < t_3 \\ (1,1) \text{ for } t_3 \le t < t_4 \\ (-1,1) \text{ for } t_4 \le t \le t_f \end{array} \right\}. \tag{97}$$

Since the initial point $(q_1(0), q_2(0), \omega_1(0), \omega_2(0))$ is specified, the optimization variable in the optimization problem (50), resp., (61) is

$$z = (\xi_1, \xi_2, \xi_3, \xi_4, \xi_5), \quad \xi_k = t_k - t_{k-1}\,, \quad k = 1, \ldots, 5.$$

The code NUDOCCCS yields the arc durations and switching times

$$\begin{array}{lll} t_1 = 0.546174\,, & \xi_2 = 1.21351\,, & t_2 = 1.75968\,, \\ \xi_3 = 1.03867\,, & t_3 = 2.79835\,, & \xi_4 = 0.906039\,, \\ t_4 = 3.70439\,, & \xi_5 = 0.185023\,, & t_f = 3.889409\,, \end{array} \tag{98}$$

as well as the initial values of the adjoint variables,

$$\begin{array}{ll} \lambda_1(0) = 0.184172\,, & \lambda_2(0) = -0.011125\,, \\ \lambda_3(0) = 1.482636\,, & \lambda_4(0) = \quad 0.997367\,. \end{array} \tag{99}$$

The strict bang-bang property (35) and the inequalities $D^k(H) > 0$, $k = 1,2,3$, hold in view of

$$\begin{array}{llll} \sigma_1(t) \neq 0 & \text{for} & t \neq t_2, t_4\,, & \dot{\sigma}_1(t_2) < 0\,, \quad \dot{\sigma}_1(t_4) > 0\,, \\ \sigma_2(t) \neq 0 & \text{for} & t \neq t_1, t_3\,, & \dot{\sigma}_2(t_1) > 0\,, \quad \dot{\sigma}_2(t_3) < 0. \end{array}$$

For the terminal conditions in (92), the Jacobian in the optimization problem is computed as the (4×5)-matrix

$$\widetilde{\mathscr{G}}_\xi(\hat{\xi}) = \begin{pmatrix} -10.858 & -12.746 & -5.8833 & -1.1500 & 0 \\ 0.19928 & -2.7105 & -1.4506 & -1.9148 & -4.83871 \\ -0.62256 & 3.3142 & 2.3155 & 2.9435 & 6.1936 \\ 9.3609 & 3.0393 & 0.48446 & 0.040581 & 0 \end{pmatrix}$$

which has full rank. The Hessian of the Lagrangian is given by

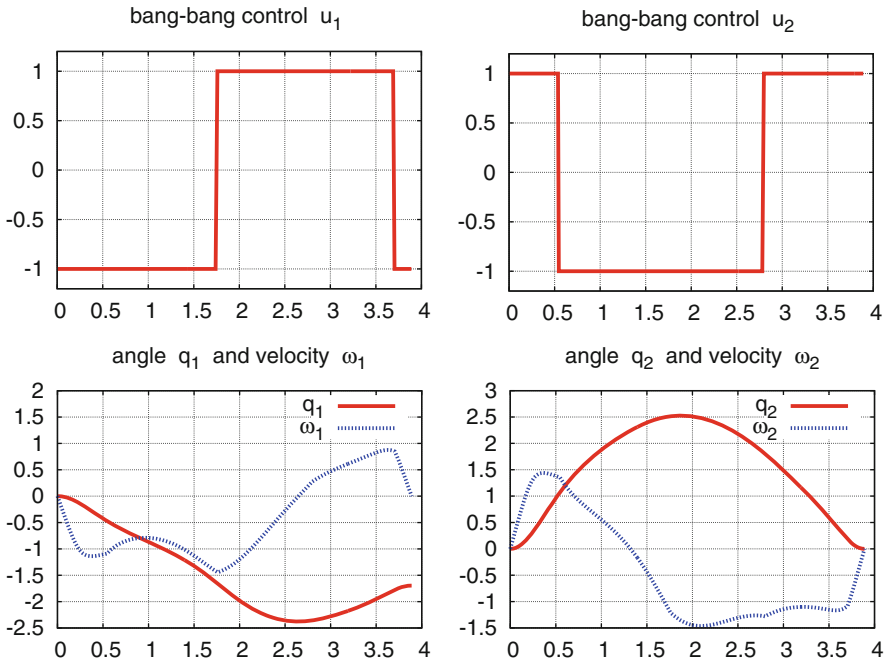

Fig. 9 Control of the two-link robot (90)–(94). *Top row*: (*left*) control u_1, (*right*) control u_2. *Bottom row*: (*left*) angle q_1 and velocity ω_1, (*right*) angle q_2 and velocity ω_2

$$\widetilde{\mathscr{L}}_{\xi\xi}(\hat{\xi},\beta) = \begin{pmatrix} 71.142 & 90.761 & 42.130 & 8.4989 & -0.051822 \\ 90.761 & 112.54 & 51.313 & 10.769 & 0.14985 \\ 42.130 & 51.313 & 23.963 & 5.1240 & 0.13860 \\ 8.4989 & 10.769 & 5.1240 & 1.4999 & 0.17078 \\ -0.051822 & 0.14985 & 0.13860 & 0.17078 & 0.29736 \end{pmatrix}.$$

This yields the projected Hessian (64) as the positive number

$$N^* \widetilde{\mathscr{L}}_{\xi\xi}(\hat{\xi},\beta) N = 0.326929\,.$$

Hence, all conditions in Theorem 17 are satisfied and thus the extremal (97)–(99) yields a strict strong minimum.

It is interesting to note that there exists a *second local minimum* with the same terminal time $t_f = 3.88941$. Though the control has also five bang-bang arcs, the control structure is substantially different from that in (97),

$$u(t) = (u_1(t), u_2(t)) = \left\{ \begin{array}{r} (1,-1)\ ,\ 0 \le t < t_1 \\ (-1,-1)\ ,\ t_1 \le t < t_2 \\ (-1,1)\ ,\ t_2 \le t < t_3 \\ (1,1)\ ,\ t_3 \le t < t_4 \\ (1,-1)\ ,\ t_4 \le t \le t_f \end{array} \right\}, \tag{100}$$

where $0 < t_1 < t_2 < t_3 < t_4 < t_5 = t_f$. NUDOCCCS determines the switching times

$$\begin{array}{l} t_1 = 0.1850163\,,\ t_2 = 1.091075\,,\ t_3 = 2.129721\,, \\ t_4 = 3.343237\,,\quad t_f = 3.889409\,, \end{array} \tag{101}$$

for which the strict bang-bang property (35) holds and $D^k(H) > 0$ for $k = 1,2,3,4$. Moreover, computations show that $rank\,(\widetilde{\mathscr{G}}_\xi(\hat{\xi})) = 4$ and that the projected Hessian of the Lagrangian (64) is the positive number

$$N^* \widetilde{\mathscr{L}}_{\xi\xi}(\hat{\xi},\beta) N = 0.326929\,.$$

It is remarkable that this value is identical with the value of the projected Hessian for the first local minimum. Therefore, also for the second solution we have verified that all conditions in Theorem 17 hold, and thus the extremal (100), (101) is a strict strong minimum. The phenomenon of multiple local solutions all with the same minimal time t_f has also been observed by Betts [3], Example 6.8 (Reorientation of a rigid body).

9 Optimal Control Problems with Mixed Control-State Constraints and Control Appearing Linearly

To the best of our knowledge, second-order sufficient optimality conditions (SSC) for optimal control problems with mixed control-state constraints have only been studied for the class of *regular controls*, where the *strict* Legendre–Clebsch condition holds. Such control problems have not yet been considered, when the control variable appears linearly in the system dynamics and in the mixed control-state constraint. For a two-sided control-state constraint we will show that the constraining function itself can be taken as a new control variable, whereby the original control problem is transformed into a classical control problem with an affine control variable subject to simple control bounds. Hence, optimal controls for the transformed control problem are concatenations of bang-bang and singular arcs. The material in this section is based on our paper [23].

9.1 Statement of the Problem and Transformed Control Problem

For simplicity, we consider an optimal control problem with fixed initial time $t_0 = 0$, fixed initial conditions and terminal equality constraints, and with a scalar control. Let $x \in \mathbb{R}^n$ denote the state variable and $u \in \mathbb{R}$ be the control variable. The terminal time $t_f > 0$ is either fixed or free. The dynamic equation and boundary conditions are

$$\dot{x} = f(t,x) + g(t,x)u, \quad x(0) = x_0, \quad K(x(t_f)) = 0. \tag{102}$$

We consider a two-sided mixed control-state constraint which is affine in the control variable:

$$\alpha \le a(x(t)) + b(x(t))\,u(t) \le \beta \quad \text{for } a.e.\ t \in [0,t_f]\,. \tag{103}$$

The optimal control problem consists in finding a control $u \in L^\infty([0,t_f],\mathbb{R})$ that *minimizes* the objective functional in Mayer form

$$\mathscr{J}(x,u) = J(x(t_f)). \tag{104}$$

The functions $f, g : \mathbb{R}^n \to \mathbb{R}^n$, $a, b : \mathbb{R}^n \to \mathbb{R}$, $J : \mathbb{R}^n \to \mathbb{R}$ and $K : \mathbb{R}^n \to \mathbb{R}^{d(K)}$ $(0 \le d(K) \le n)$ are assumed to be twice continuously differentiable. We remind the reader that a Bolza functional of the form

$$\mathscr{J}(x,u) = J(x(t_f)) + \int_0^{t_f} (f_0(t,x) + g_0(t,x)u)dt \tag{105}$$

can be reduced to Mayer form by introducing the additional state variable y that solves the initial value problem $\dot{y} = f_0(t,x) + g_0(t,x)u$, $y(0) = 0$ and minimizing the functional $J(x(t_f)) + y(t_f)$.

The following *regularity assumption* will be assumed to hold for feasible trajectories:

$$b(t,x(t)) \neq 0 \quad \text{for} \quad t \in [0,t_f]\,. \tag{106}$$

This assumption allows us to introduce a *new control variable* v that is related to the control variable u as follows:

$$v := a(x) + b(x)\,u, \quad \text{i.e.,} \quad u = (v - a(x))/b(x). \tag{107}$$

The *transformed optimal control problem* consists in minimizing the objective (104) subject to the transformed dynamics

$$\dot{x} = \bar{f}(t,x) + \bar{g}(t,x)v, \quad x(0) = x_0, \quad K(x(t_f)) = 0, \tag{108}$$

where the transformed functions $\bar{f}, \bar{g}$ are defined by

$$\bar{f}(t,x) = f(t,x) - g(t,x)a(x)/b(x), \quad \bar{g}(t,x) = g(t,x)/b(x). \tag{109}$$

The mixed control-state constraint (103) then is equivalent to the simple control constraint

$$\alpha \leq v(t) \leq \beta. \tag{110}$$

Thus, we can apply the second-order conditions developed in Sects. 5 and 6 to the transformed problem.

9.2 Numerical Example: Optimal Control of the Rayleigh Equation

The Rayleigh equation describes oscillations of the electric current, resp., voltage in an electric circuit. The optimal control of the Rayleigh equation for a control-quadratic objective has been studied in Maurer and Augustin [20], Osmolovskii and Maurer [36], and Chen and Gerdts [8], where both simple control bounds and a mixed control-state constraint were investigated.

Let x_1 denote the electric current and x_2 the voltage. The control u represents the voltage at the generator which steers the following dynamic equations:

$$\begin{aligned} \dot{x}_1 &= x_2, & x_1(0) &= -5, \\ \dot{x}_2 &= -x_1 + x_2(1.4 - 0.14x_2^2) + u, & x_2(0) &= -5. \end{aligned} \tag{111}$$

We consider the mixed control-state constraint

$$\alpha \le u + x_1 \le \beta \quad (\alpha = -5,\ \beta = 0). \tag{112}$$

Various other bounds α and β have been studied in Maurer and Omolovskii [23]. The mixed constraint is a slight modification of the one considered in [8, 20]. The objective is to minimize the quadratic functional

$$\mathscr{J}(x,u) = \int_0^{t_f} (x_1(t)^2 + x_2(t)^2)dt\,. \tag{113}$$

First, we consider this control problem with fixed terminal time $t_f = 4.5$. Later, we shall prescribe the terminal condition

$$x_1(t_f) = 0 \tag{114}$$

and solve the control problem with free terminal time t_f.

According to (107), the new control variable v is given by

$$v = u + x_1\,, \quad u = v - x_1\,. \tag{115}$$

The Pontryagin function (Hamiltonian) with respect to the control v becomes

$$H(x,\lambda,v) = x_1^2 + x_2^2 + \lambda_1 x_2 + \lambda_2(-x_1 + x_2(1.4 - 0.14x_2^2) + v - x_1). \tag{116}$$

The adjoint equations are

$$\begin{aligned} \dot\lambda_1 &= -H_{x_1} = -2x_1 + 2\lambda_2, \\ \dot\lambda_2 &= -H_{x_2} = -2x_2 - \lambda_1 + \lambda_2(0.42x_2^2 - 1.4). \end{aligned} \tag{117}$$

For the first control problem with free endpoint $x(t_f)$ and fixed final time $t_f = 4.5$, we get the transversality condition

$$\lambda_1(t_f) = 0,\ \lambda_2(t_f) = 0, \tag{118}$$

while the second control problem with terminal constraint $x_1(t_f) = 0$ and free terminal time t_f gives the transversality conditions

$$H(t_f) = 0,\ \lambda_2(t_f) = 0. \tag{119}$$

The *switching function* $\sigma = H_v = \lambda_2$ determines the optimal control according to

$$v(t) = \begin{cases} \beta & , \text{ if } \quad \lambda_2(t) < 0, \\ \alpha & , \text{ if } \quad \lambda_2(t) > 0, \\ \text{singular} & , \text{ if } \quad \lambda_2(t) = 0 \quad \forall\, t \in I_s \subset [0, t_f]. \end{cases} \tag{120}$$

We do not discuss singular controls further, because for the chosen bounds $\alpha = -5$ and $\beta = 0$ in the mixed constraint (112) we obtain bang-bang controls. Singular controls for smaller values of α are discussed in [23].

In the first control problem with free terminal state $x(t_f)$ and fixed terminal time t_f, we obtain a bang-bang control with three bang-bang arcs:

$$v(t) = \left\{ \begin{array}{lll} 0 & , \text{ if } & 0 \le t \le t_1 \\ -5 & , \text{ if } & t_1 < t \le t_2 \\ 0 & , \text{ if } & t_2 < t \le t_f \end{array} \right\}. \tag{121}$$

The code NUDOCCCS [4] yields the following results:

$$\begin{array}{lll} \mathscr{J}(x,u) = 62.165171, & t_1 = 0.77996717, & t_2 = 2.6835574, \\ x_1(t_f) = -0.40342897, & x_2(t_f) = -1.4332277, & \\ \lambda_1(0) = -13.364385, & \lambda_2(0) = -5.591549. & \end{array}$$

The corresponding extremal solution is shown in Fig. 10.

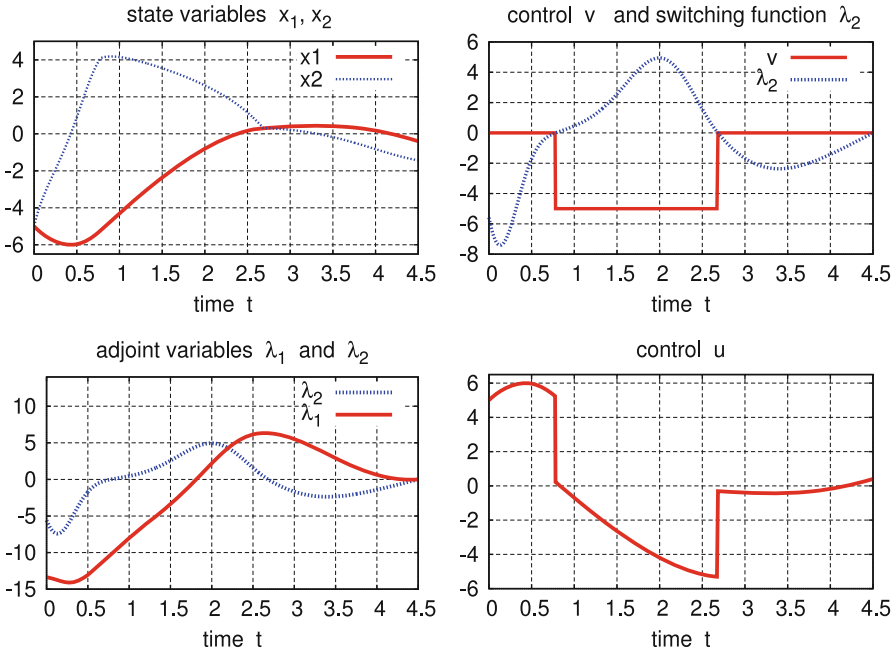

Fig. 10 Objective (113) and constraint $-5 \le v = u + x_1 \le 0$. *Top row*: (*left*) state variables x_1, x_2, (*right*) transformed control v and switching function $\sigma = \lambda_2$. *Bottom row*: (*left*) adjoint variables λ_1, λ_2, (*right*) control u

Next, we compute the Hessian of the Lagrangian for the IOP:

$$\tilde{\mathscr{L}}_{\tilde{z}\tilde{z}} = \begin{pmatrix} 573.237 & 458.854 \\ 458.854 & 377.399 \end{pmatrix}.$$

Obviously, this matrix is positive-definite. Moreover, Fig. 10 (top row, right)shows that the switching function $\sigma(t) = \lambda_2(t)$ satisfies the strict bang-bang property (37); cf. also Remark 5.1:

$$\begin{array}{llll} \lambda_2(t) < 0 & \forall\ 0 \le t < t_1, & \lambda_2(t_1) = 0, & \dot{\lambda}_2(t_1) > 0, \\ \lambda_2(t) > 0 & \forall\ t_1 < t < t_2, & \lambda_2(t_2) = 0, & \dot{\lambda}_2(t_2) < 0, \\ \lambda_2(t) < 0 & \forall\ t_2 < t < t_f, & \lambda_2(t_f) = 0, & \dot{\lambda}_2(t_f) > 0. \end{array}$$

Hence, the extremal shown in Fig. 10 satisfies the SSC in Theorem 17 and thus is a strict strong minimum.

Now we study the solution, when the terminal condition $x_1(t_f) = 0$ is imposed and the terminal time t_f is free. In this case we obtain a bang-bang control with only one switch:

$$v(t) = \left\{ \begin{array}{ll} 0 \ , \text{ if} & 0 \le t \le t_1 \\ -5 \ , \text{ if} & t_1 < t \le t_f \end{array} \right\}. \tag{122}$$

The corresponding IOP has the two optimization variables t_1, t_f and the scalar equality constraint $x_1(t_f) = 0$. The code NUDOCCCS yields the following results:

$$\begin{array}{lll} \mathscr{J}(x,u) = 60.72697, & t_1 = 0.8343100, & t_f = 2.364688, \\ x_1(t_f) = 0.0, & x_2(t_f) = 1.600202, & \\ \lambda_1(0) = -13.4868, & \lambda_1(t_f) = -1.60020, & \\ \lambda_2(0) = -5.72868, & \lambda_2(t_f) = 0.0. & \end{array}$$

The extremal solution with state, control, adjoint variables, and switching function is shown in Fig. 11.

The reduced Hessian (64) of the Lagrangian for the IOP is a scalar which we compute as the positive number 4.3525. Moreover, Fig. 11, top row, right, shows that the strict bang-bang property is fulfilled; cf. also Remark 5.1:

$$\begin{array}{llll} \lambda_2(t) < 0 & \forall\ 0 \le t < t_1, & \lambda_2(t_1) = 0, & \dot{\lambda}_2(t_1) > 0, \\ \lambda_2(t) > 0 & \forall\ t_1 < t < t_f, & \lambda_2(t_f) = 0, & \dot{\lambda}_2(t_f) < 0. \end{array}$$

Hence, the extremal shown in Fig. 11 satisfies the SSC in Theorem 17 and thus provides a strict strong minimum.

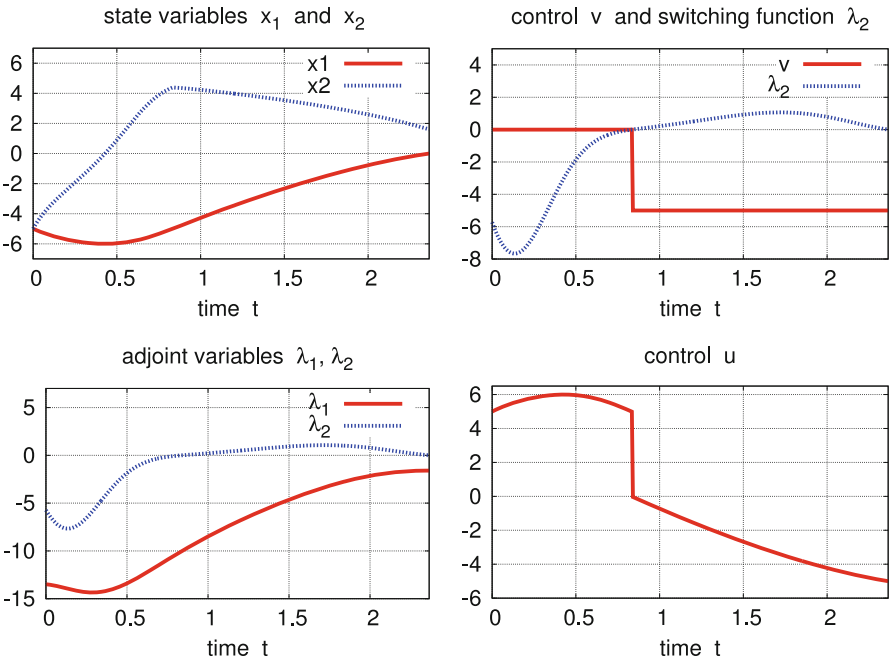

Fig. 11 Objective (113) with free terminal time t_f and constraints $-5 \le v = u + x_1 \le 0, x_1(t_f) = 0$. *Top row*: (*left*) state variables x_1, x_2, (*right*) transformed control v and switching function $\sigma = \lambda_2$. *Bottom row*: (*left*) adjoint variables λ_1, λ_2, (*right*) control $u = v - x_1$

10 Conclusion

We presented no-gap necessary and sufficient second-order optimality conditions for extremals with discontinuous controls in the simplest problem of the Calculus of Variations and the general optimal control problem with regular mixed constraint $g(t,x,u) = 0$ on a variable time interval $[t_0, t_f]$. We formulated similar conditions for bang-bang controls in an optimal control problem with a Mayer functional, where the dynamical system is affine in control variable and the control constraint is given by a convex polyhedron. Bang-bang controls induce an optimization problem with respect to the switching times of the control, the so-called Induced Optimization Problem IOP. We showed that the classical second-order sufficient condition for the IOP together with the strict bang-bang property of the switching function ensure second-order sufficient conditions (IOP) for the bang-bang control problem. The verification of SSC for bang-bang controls was illustrated on two numerical examples. First, we studied extremals in the optimal control of the chemotherapy of HIV. Then, following [36], we investigated extremals in time-optimal control problems of two-link robots. We also discussed optimal control problems with running mixed control-state constraints and control appearing linearly. Taking the

mixed constraint as a new control variable we converted such problems to bang-bang control problems. As an example, we studied extremals in the optimal control problem for the Rayleigh equation.

The results on SSC naturally lend themselves to sensitivity results for the IOP and the underlying bang-bang control problem using the well-known sensitivity results for finite-dimensional optimization problems developed by Fiacco [11]. Sensitivity results for bang-bang controls may be found in Felgenhauer [10], Kim and Maurer [17] and Maurer and Vossen [25]. Sensitivity results also allow to develop real-time control techniques as indicated already in Büskens et al. [7]. These issues will be the topic of a future paper. Related sensitivity results may be obtained for bang-singular controls as suggested in Vossen [40, 41]. This approach still needs a practical method of verifying the more abstract SSC for bang-singular controls given in Aronna et al. [2].

References

1. Agrachev, A.A., Stefani, G., Zezza, P.L.: Strong optimality for a bang-bang trajectory. SIAM J. Control Optim. **41**, 991–1014 (2002)
2. Aronna, M.S., Bonnans, F., Dmitruk, A., Lotito, P.A.: Quadratic order conditions for bang-singular extremals. Numer. Algebra Control Optim. **2**, 511–546 (2012)
3. Betts, J.T.: Practical Methods for Optimal Control and Estimation Using Nonlinear Programming. Advances in Design and Control, vol. DC 19, 2nd edn. SIAM Publications, Philadelphia (2010)
4. Büskens, C.: Optimierungsmethoden und Sensitivitätsanalyse für optimale Steuerprozesse mit Steuer- und Zustands-Beschränkungen. Dissertation, Institut für Numerische Mathematik, Universität Münster (1998)
5. Büskens, C., Maurer, H.: SQP-methods for solving optimal control problems with control and state constraints: adjoint variables, sensitivity analysis and real-time control. J. Comput. Appl. Math. **120**, 85–108 (2000)
6. Büskens, C., Maurer, H.: Sensitivity analysis and real-time optimization of parametric nonlinear programming problems. In: Grötschel, M., Krumke, S.O., Rambau, J. (eds.) Online Optimization of Large Scale Systems, pp. 3–16. Springer, Berlin (2001)
7. Büskens, C., Pesch, H.J., Winderl, S.: Real-time solutions of bang-bang and singular optimal control problems. In: Grötschel, M., Krumke, S.O., Rambau, J. (eds.) Online Optimization of Large Scale Systems, pp. 129–142. Springer, Berlin (2001)
8. Chen, J., Gerdts, M.: Smoothing techniques of nonsmooth Newton methods for control-state constrained optimal control problems. SIAM J. Numer. Anal. **50**, 1982–2011 (2012)
9. Chernousko, F.L., Akulenko, L.D., Bolotnik, N.N.: Time-optimal control for robotic manipulators. Opt. Control Appl. Methods **10**, 293–311 (1989)
10. Felgenhauer, U.: On stability of bang-bang type controls. SIAM J. Control Optim. **41**, 1843–1867 (2003)
11. Fiacco, A.V.: Introduction to Sensitivity and Stability Analysis in Nonlinear Programming. Mathematics in Science and Engineering, vol. 165. Academic, New York (1983)
12. Fourer, R., Gay, D.M., Kernighan, B.W.: AMPL: A Modeling Language for Mathematical Programming. Duxbury Press/Brooks-Cole Publishing Company, Monterey (1993)
13. Geering, H.P., Guzella, L., Hepner, S.A.R., Onder, C.: Time-optimal motions of robots in assembly tasks. IEEE Trans. Autom. Control **AC-31**, 512–518 (1986)

14. Göllmann, L.: Numerische Berechnung zeitoptimaler Trajektorien für zweigliedrige Roboterarme. Diploma thesis, Institut für Numerische Mathematik, Universität Münster (1991)
15. Hannemann, R.: Diploma thesis, Institut für Numerische und Angewandte Mathematik, Universität Münster (2004)
16. Kaya, C.Y., Noakes, J.L.: Computational method for time-optimal switching control. J. Optim. Theory Appl. **117**, 69–92 (2003)
17. Kim, J.-H.R., Maurer, H.: Sensitivity analysis of optimal control problems with bang-bang controls. In: Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, pp. 3281–3286 (2003)
18. Kirschner, D., Lenhart, S., Serbin, S.: Optimal control of the chemotherapy of HIV. J. Math. Biol. **35**, 775–792 (1996)
19. Malanowski, K., Maurer, H.: Sensitivity analysis for parametric control problems with control-state constraints. Comput. Optim. Appl. **5**, 253–283 (1996)
20. Maurer, H., Augustin, D.: Sensitivity analysis and real-time control of parametric optimal control problems using boundary value methods. In: Grötschel, M., Krumke, S.O., Rambau, J. (eds.) Online Optimization of Large Scale Systems, pp. 17–55. Springer, Berlin (2001)
21. Maurer, H., Osmolovskii, N.P.: Quadratic sufficient optimality conditions for bang-bang control problems. Control Cybern. **33**, 555–584 (2003)
22. Maurer, H., Osmolovskii, N.P.: Second order sufficient conditions for time-optimal bang-bang control problems. SIAM J. Control Optim. **42**, 2239–2263 (2004)
23. Maurer, H., Osmolovskii, N.: Second-order conditions for optimal control problems with mixed control-state constraints and control appearing linearly. In: Proceedings of the 52nd IEEE Conference on Control and Design, Firenze, 9–12 Dec, pp. 514–519 (2013)
24. Maurer, H., Pickenhain, S.: Second order sufficient conditions for optimal control problems with mixed control-state constraints. J. Optim. Theory Appl. **86**, 649–667 (1995)
25. Maurer, H., Vossen, G.: Sufficient conditions and sensitivity analysis for bang-bang control problems with state constraints. In: Korytowski, A., Szymkat, M. (eds.) Proceedings of the 23rd IFIP Conference on System Modeling and Optimization, Cracow, pp. 82–99. Springer, Berlin (2009)
26. Maurer, H., Büskens, C., Kim, J.-H.R., Kaya, Y.: Optimization methods for the verification of second-order sufficient conditions for bang-bang controls. Opt. Control Methods Appl. **26**, 129–156 (2005)
27. Maurer, H., Tarnopolskaya, T., Fulton, N.: Computation of optimal bang-bang and singular controls in planar collision avoidance. J. Ind. Manag. Optim. **10**(2), 443–460 (2014) [Special Issue on Computational Methods for Optimization and Control]
28. Milyutin, A.A., Osmolovskii, N.P.: Calculus of Variations and Optimal Control. Translations of Mathematical Monographs, vol. 180, American Mathematical Society, Providence, RI (1998)
29. Noble, J., Schättler, H.: Sufficient conditions for relative minima of broken extremals. J. Math. Anal. Appl. **269**, 98–128 (2002)
30. Oberle, H.J.: Numerical computation of singular control functions for a two-link robot arm. Opt. Control Lect. Notes Control Inf. Sci. **95**, 244–253 (1987)
31. Osmolovskii, N.P.: Quadratic optimality conditions for broken extremals in the general problem of calculus of variations. J. Math. Sci. **123**(3), 3987–4122 (2004)
32. Osmolovskii, N.P.: Second order conditions in optimal control problem with mixed equality-type constraints on a variable time interval. Control Cybern. **38**(4) 1535–1556 (2009)
33. Osmolovskii, N.P.: Sufficient quadratic conditions of extremum for discontinuous controls in optimal control problem with mixed constraints. J. Math. Sci. **173**, 1–106 (2011)
34. Osmolovskii, N.P.: Necessary quadratic conditions of extremum for discontinuous controls in optimal control problem with mixed constraints. J. Math. Sci. **183**, 435–576 (2012)
35. Osmolovskii, N.P., Maurer, H.: Equivalence of second order optimality conditions for bang-bang control problems. Part 2: Proof, variational derivatives and representations. Control and Cybern. **36**, 5–45 (2007); Part 1: Main result. Control Cybern. **34**, 927–950 (2005)

36. Osmolovskii, N.P., Maurer, H.: Applications to Regular and Bang-Bang Control: Second-Order Necessary and Sufficient Optimality Conditions in Calculus of Variations and Optimal Control. SIAM Series Design and Control, vol. DC 24. SIAM Publications, Philadelphia (2012)
37. Perelson, A., Kirschner, D., DeBoer, R.: The dynamics of HIV infection of $CD4^+$ T cells. Math. Biosci. **114**, 81–125 (1993)
38. Schättler, H., Ledzewicz, U.: Geometric Optimal Control: Theory, Methods and Examples. Interdisciplinary Applied Mathematics, vol. 38. Springer, New York (2012)
39. Schättler, H., Ledzewicz, U., Maurer, H.: Sufficient conditions for strong local optimality in optimal control problems with L_2-type objectives and control constraints. Discrete Contin. Dyn. Syst. B **19**(8), 2657–2679 (2014)
40. Vossen, G.: Numerische Lösungsmethoden, hinreichende Optimalitätsbedingungen und Sensitivitätsanalyse für optimale bang-bang und singuläre Steuerungen. Dissertation, Institut für Numerische und Angewandte Mathematik, Universität Münster (2005)
41. Vossen, G.: Switching time optimization for bang-bang and singular controls. J. Optim. Theory Appl. **144**(2), 409–429 (2010)
42. Wächter, A., Biegler, L.T.: On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Math. Program. **106**, 25–57 (2006); cf. IPOPT home page (C. Laird and A. Wächter): https://projects.coin-or.org/Ipopt